AAPM REPORT NO. 43

QUALITY ASSESSMENT AND IMPROVEMENT OF DOSE RESPONSE MODELS: SOME EFFECTS OF STUDY WEAKNESSES ON STUDY FINDINGS. "C'EST MAGNIFIQUE?"

Published for the
American Association of Physicists in Medicine
by the American Institute of Physics

AAPM REPORT NO. 43

QUALITY ASSESSMENT AND IMPROVEMENT OF DOSE RESPONSE MODELS: SOME EFFECTS OF STUDY WEAKNESSES ON STUDY FINDINGS. "C'EST MAGNIFIQUE ?"

A Report of Task Group 1, "Evaluation of Models for Dose-Response
in Radiation Oncology," of the Biological Effects Committee
American Association of Physicists in Medicine

15 June 1993

Published for the
American Association of Physicists in Medicine
by Medical Physics Publishing Corp.

Disclaimer: This publication is based on sources and information believed to be reliable, but the AAPM and the editors disclaim any warranty or liability based on or relating to the contents of this publication.

The AAPM does not endorse any products, manufacturers or suppliers. Nothing in this publication should be interpreted as implying such endorsement.

Further copies of this report may be obtained from:
American Association of Physicists in Medicine
One Physics Ellipse
College Park, MD 20740-3846

International Standard Book Number: 0-944838-45-6

Published by Medical Physics Publishing Corporation
732 N. Midvale Boulevard
Madison, WI 53705
(608) 262-4021, FAX: (608) 265-2121

Printed in the United States of America

AAPM REPORT NO. 43

QUALITY ASSESSMENT AND IMPROVEMENT OF DOSE RESPONSE MODELS: SOME EFFECTS OF STUDY WEAKNESSES ON STUDY FINDINGS. "C'EST MAGNIFIQUE ?"

A Report of Task Group 1, "Evaluation of Models for Dose-Response
in Radiation Oncology," of the Biological Effects Committee
American Association of Physicists in Medicine

Donald E. Herbert, Ph.D., principal author.
Chairman, Task Group 1, 1984-1992
Chairman, Biological Effects Committee, 1985-1991

Timothy Schultheiss, Ph.D.
Chairman, Biological Effects Committee, 1992-present

Arnold Feldman, Ph.D.
Engikolai Krishnan, M.D.
Colin Orton, Ph.D.
Jacques Ovadia, Ph.D.
Bhudatt Paliwal, Ph.D.
Timothy Schultheiss, Ph.D.
Prakash Shrivastava, Ph.D.
Alfred Smith, Ph.D.
Marilyn Stovall, MPH
Lionel Cohen, M.D., Ph.D. (Consultant)

15 June 1993

Published for the
American Association of Physicists in Medicine
by Medical Physics Publishing Corp.

"Even Jehovah

After Moses had got the Commandments

Committed to stone

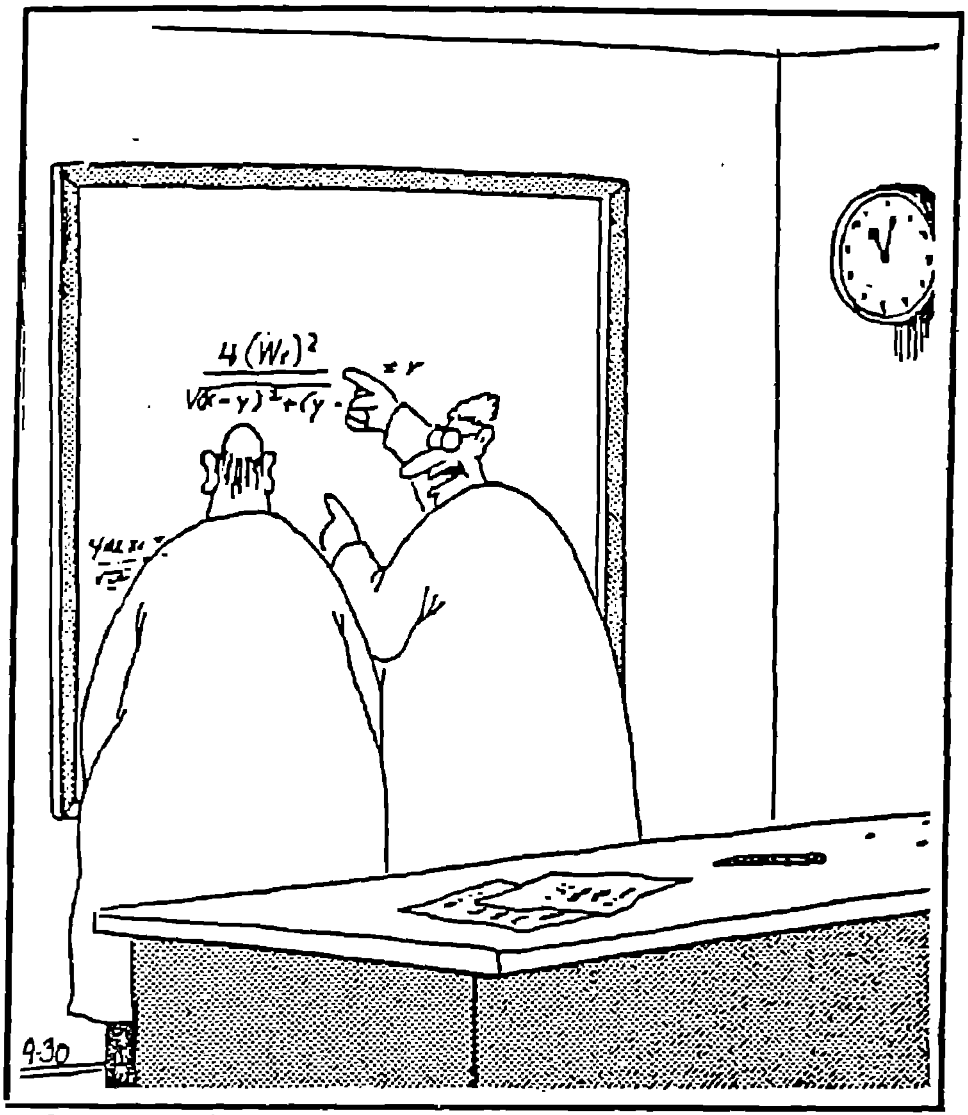
Probably thought:

I always forget the things

I really intended to say."

C. Morley, 1923

THE FAR SIDE GARY LARSON



"Yes, yes, I know that Sydney ... everybody knows that! But look: Four wrongs SQUARED, minus two wrongs to the fourth power, divided by this formula, DO make a right."

PREFACE

Practicing radiation oncologists may not be aware of a crisis in the theoretical underpinnings of their profession. In this Monograph a "task-group" of the AAPM, with the continued support and guidance of Professor Donald Herbert, who first noted insistent problems arising in the interpretation and application of radiobiological data, reports on the statistical inadequacy of many historical and contemporary studies which have been accepted as a basis for the construction of models used in clinical practice. Use of these flawed models may well have retarded progress in the past and, more importantly, still threaten to inhibit future advances in the practice of radiation oncology.

Why were such models developed? Is there any need for modelling in clinical practice? The rationale is roughly as follows. It has long been known that many forms of ionizing radiation could kill cells, or at least arrest cell division, so as to disrupt the structure or function of organs and tissues. Of particular interest was the capacity to interrupt and reverse the growth of malignant tumors. With adequate dosage, and a judicious choice of technical factors, this effect could lead to a complete and permanent cure of the disease without serious damage to normal tissues. This observation, with its extraordinary promise as an effective cancer remedy, was intensively studied in both clinic and laboratory. Nine decades of clinical practice provided a vast amount of empirical data on the response of human tissues and tumors to ionizing radiation. Many thousands of patients had been irradiated using whatever combinations of technical factors (choice of particle, energy, dosage, target volume, fractionation and time) that seemed appropriate and practical at the time. Some combinations proved safer and more effective than others, and these have become accepted standards of practice in clinical cancer therapy. The choice of particle (photons, electrons or neutrons, for example), the energy required for adequate penetration, and the target or tumor volume to be irradiated, are usually determined by physical and anatomical constraints not under the control of the prescribing radiotherapist. On the other hand, the dose delivered, the fraction size and number, and the total treatment time (or the average interval between fractions given and the number of fractions), are all readily controllable independent variables. The practicing radiation oncologist perceived it as a duty to find that particular combination of technical factors which would predictably lead to a desired clinical outcome.

The effect of varying the dose was first studied. It would have been useful to be able to identify a "threshold", or minimal effective dose, that would lead to a specific response in a given patient. However, unlike other therapeutic agents (pharmaceuticals, for example) it is not possible, because of the cumulative effects of radiation, to "titrate" radiation dosage so as to determine an individual threshold explicitly. If a particular dose yielded an observed quantal (all-or-nothing) response, there was no way to tell whether this dose was just at, or well above, the threshold value. The best that could be done was to observe the responses in many individuals exposed to various dose levels, and then determine the frequency (or probability) of a particular predefined end-result (tumor cure or normal tissue damage) as a function of dose (for constant field-size and fractionation). This procedure usually gave a characteristically sigmoid dose-response curve. The dose-response function could be modelled (and quantitated) if the individual threshold doses were assumed to be normally distributed, with parameters defining the mean value and standard deviation of the distribution. Then the corresponding cumulative frequency integral (probit or logit) generates an analogous sigmoid dose-response function with equivalent parameters for position and slope. The two parameters are likely to be tissue-specific and measurably different for normal tissue and tumor. This simple model has some significant consequences.

It was realized that if the respective dose-response parameters for both a given human normal tissue and an associated tumor were known, then a "therapeutic ratio" (ratio of median response doses) could be derived explicitly. Similarly, a "prognosis" or probability of clinical success - that is the conditional probability of tumor control without normal tissue injury - could be calculated precisely for any given dose. With two sigmoid dose-response curves for control and complications respectively, the probability of uncomplicated control is always a monophasic, bell-shaped dose-response function with a single, well-defined maximum corresponding to an "optimal" dose. Paterson and others had presented some clinical evidence for the existence of an optimal dose, yielding the best chance of uncomplicated cure, and measurably lower success rates at all doses above and below this critical value. These observations had important implications for clinical radiation oncology and provided a powerful incentive to validate the dose-response model and derive accurate estimates of the relevant parameters.

Up to this point the system is reasonably tractable. With some little effort, the required parameters could be estimated and a model with considerable predictive capacity would exist. However, this possibility is severely constrained by uncertainties in dose determination, due to several confounding factors, most notably the heterogeneity of absorbed dose (dose is not a single number but a gradient varying substantially throughout the target volume), time factors (effects of repair and repopulation), and volume effects. Including these variables makes the already complex optimization process enormously more complicated.

Historically, the next important study concerned the effect of time factors and the dose-time relationship. It was noted that when treatment was fractionated and the total treatment time more protracted, larger doses were generally required to achieve an equivalent response. This effect was attributed to "recovery" from the induced radiation damage (At this stage the separate effects of fractionation and protraction, or repair and repopulation, were not recognized). In 1944 Strandqvist suggested a relatively simple dose-time relationship for equivalent effects, namely a linear dependence of log-dose upon log-time. This was an iso-effect function, quite distinct from the dose-response functions described previously. Again, since the observed effects were quantal, apparently similar reactions did not necessarily imply that the doses were biologically equivalent. When a given dose-time combination yielded a particular observed reaction, the dose may have been just at, or well above, the threshold value. Similarly, when the treatment failed to produce the expected reaction, the dose may have been just below, or far smaller than, the threshold value. Under the circumstances, it was not possible to fit an iso-effect line to the data by conventional regression (least squares) techniques.

Strandqvist adopted the stratagem of plotting individual observations relating dose to overall time on a log-log graph and then fitting (by eye) a single straight line separating, as far as possible, responders above the line from non-responders below. This line was believed to indicate equivalent or "iso-effective" doses for all possible treatment times, and had two parameters: position and slope. This type of fitting process is highly subjective. If the two groups were well separated, any number of lines could be drawn between them; position and slope would then be indeterminate. In Standqvist's series the two groups overlapped substantially. In principle, given enough data and using modern statistical methods, a best-fitting iso-response line could have been identified. In the event, neither the size of the data set nor the statistical procedures then available, were adequate for this purpose. Standqvist could not provide confidence limits on his crude estimates of position and slope, nor could he discern differences, if any, between the tumor and associated normal tissues.

Subsequent attempts to simulate iso-effective dose-time relationships, using either empirical power functions (NSD or TDF) or cell population kinetic models (multi-target or linear-quadratic), have not yet yielded any strikingly better methods for optimizing the clinical outcome. The major deficiency in all the proposed models is the difficulty in deriving accurate tissue-specific parameters. Our belief is that a substantial improvement in treatment planning and optimization would become possible IF the model used were realistic, or at least plausible, IF the relevant parameters could be derived from available data with confidence, or at least with known confidence limits, IF the parameters derived for normal tissue and tumor response were measurably different, and IF the associated dose-response vectors could be extracted from the analysis so as to estimate the dose, for any given treatment time, required to reach a particular response level (say, 5% risk of injury or 95% tumor control).

The main object of the task-group was to review and re-evaluate available models in the light of the data upon which they were constructed as well as any more reliable data which may have become available. The group utilized modern statistical methods for defining the goodness of fit of various alternative models to the data, and for deriving best-fitting parameters and their confidence limits for specific tissues and tumors. The information derived would create the possibility of individualizing and optimizing treatment for any specific tumor type in a given location and selecting, from among all possible combinations of dose, time and volume factors, that giving the best prospect of cure and a minimal risk of complications. The value of this exercise to the clinical radiation oncologist is incalculable.

Lionel Cohen, M.D., Ph.D.

Quality Assessment and Improvement for Dose Response Models.
Some Effects of Study Weaknesses on Study Findings.
"C'est Magnifique ?"

Précis

This report provides a panoramic view of the current states of the arts of the construction and deployment of radiation dose-response models. This view is based on the results of a critical evaluation by the AAPM Biological Effects Committee Task Group 1 (TG1) of the empirical evidence and theoretical arguments - concepts, data, methods and criteria - that have been presented in 40 studies of the several received models of radiation dose-response for the following end-points: radiation toxicity, mutagenesis, tumorigenesis, and lethality. These studies have been published in the peer-reviewed literature up to 1992 and have been repeatedly cited in authoritative reviews. This evaluation was based upon secondary analyses of the data from which the received models were constructed. These latter analyses were implemented by the concepts, methods, and criteria of statistical modelling: the generalized linear model, regression diagnostics, (including jackknife estimation and validation), Bayesian regression, etc.

This panoramic view is not an altogether pretty picture. But it is becoming a clear one. It is becoming evident that, on several insistent issues in radiation biology, serious debate may have been prematurely foreclosed by the peer-group. Our secondary analyses have disclosed the presence of consequential weaknesses both in the received praxis and in the models deployed in important sections of the radiobiological sciences. There are both ontological and epistemological deficiencies (ontology, - "the way the world is"; epistemology - "how we find out about it."). For example, on the ontological questions concerning the shape of dose-response curves, these analyses suggest that the inferences from the received models concerning the presence or absence in a given set of observations of such fundamental ontological categories as "shoulders", "thresholds", dose-dependence of the neutron RBE, etc., are not always substantiated and in several specific instances are simply incorrect; in particular they cannot be simply assumed to be correct in any given case. The validity of such inferences must always be tested - by statistically adequate measures. For another example, on the important epistemological issue of the concordance of a given model with a given set of data, these analyses disclosed that none of the received models "fit" a common set of data as well as did a cognate rival model, on the evidence of statistically adequate methods and criteria, although on the basis of received methods and criteria it had been reported in the literature that the received models "fit" these data. Indeed, the TG1 report is pre-occupied throughout with the fundamental epistemological issues of "How we learn from data - or fail to."

On the still more insistent issue of the conjunction of the fundamental thrust of the science of clinical radiobiological modelling with that of the craft of radiation therapy - which provides the practical justification for, or even legitimates, much of the work of the former - our reviews suggest that some redirection of effort may be appropriate for this section of the radiobiology peer-group: The professional aim of radiation therapy is to achieve the maximum probability of uncomplicated control of disease. A useful description of the problem requires a model of the joint probability of the concomitant occurrence of (at least) two quantal responses in the target volume - one in the tumor and one in the normal tissues. That is, it requires a dose-response model. However, most previous as well as most current efforts in radiobiological modelling are directed to the unevaluated (vide supra) construction of models of isoeffect for a single quantal response. But it is, of course, quite impossible to obtain from any isoeffect model those estimates of joint conditional probability that are required to inform the professional thrust of the radiation therapy peer-group.

Moreover, although the probability of occurrence of a given quantal response at a given level of total dose (D) is strongly modulated by the concomitant levels of fractionation (N) and protraction (T) adequate estimates of the respective roles of these two covariates, which cannot, of course, be readily obtained (only by non-least squares methods) from non-experimental (clinical) dose-response data because of the high correlations in the distribution of the treatment variables (dose, time, fractions, etc.) that are imposed by the "standard of practice" constraint, also cannot be obtained from current experimental data because of the presence of an equally high degree of multicollinearity (as well as other weaknesses) in the experimental designs.

Finally, the most characteristic feature of radiobiological models is Uncertainty: uncertainty as to the form of the model; uncertainty as to the values of the parameters of the model; uncertainty in the functions of the parameters of the model - linear functions such as the predicted response and non-linear functions such as a ratio of two parameters. (Indeed, in the well-known remarks of the nineteenth century physicist/philosopher, Charles Saunders Peirce, in radiobiology - as in most other important enterprises - "Chance itself pours in at every avenue of sense; it is of all things the most obtrusive.") However, although each of these uncertainties can be rigorously treated and their effects circumscribed by proper statistical methods (methods that provide the tools for, in Ian Hacking's apt phrase, (1993) "the taming of chance") none ever is in most of the radiobiological literature that we reviewed.

Thus, if it were to be found that most radiotherapists currently make little use of radiobiological models in their practice it should perhaps be no great surprise, since the current models are not much answering to their professional needs. We note here that, partly as a result of the findings presented in this report, TG1 has proposed a survey of radiation oncologists to assess the nature and degree of the deployment of radiobiological models in clinical practice.

We remark that the received dose-response - and isoeffect - models (e.g., the so-called linear-quadratic models) may be "correct" - although the latter may often be misdirected, as noted. However, empirical evidence for that proposition, based on the published findings of those studies cited in support that we have examined does not support it. (Perhaps other studies, also cited as empirical evidence for the received models but that we have not yet reviewed, will be found to support them. It must be remembered that, "absence of evidence is not necessarily evidence of absence".) The empirical evidence for those models that we have evaluated is often of such poor quality as to suggest that the model must be believed in order to be seen, suggesting further, at the least, that the received models are not "law-like". Therefore, they should not be, as they often are now, uncritically recommended for deployment in either the clinic or the laboratory. It is the case, of course, that "... acceptance of an inadequate explanation discourages search for a good one" (H. Jeffreys, 1957).

If, as many believe, "Science is the collection of reproducible facts", (Chalmers, 1987) then it is important to note that in most of those studies that we analyzed, the reported "facts" did not reproduce when examined by statistically adequate methods. (We encountered re-iterated assumptions more often than reproducible facts.) We recommend, therefore, that i) the concordance of any model of dose-response - or isoeffect - either the received or the rival, with any set of data, ii) the precision of the sample estimates, both of model parameters and of functions of model parameters, such as the predicted response derived therefrom, as well as iii) the consistency, with a priori information, of the sample estimates of the parameters and functions of the parameters, be assessed by statistically adequate methods before the estimates and inferences made from any model are deployed in any consequential clinical and laboratory enterprises.

Our analyses strongly suggest that much of what is published in the field of modelling of radiation dose-response - and isoeffect - may be properly termed, in the pathologist/philosopher Ludwig Fleck's locution, exoteric - as distinct from esoteric - science: "Simplified, lucid, and apodictic science - these are the most important characteristics of exoteric knowledge. In place of the specific constraint of thought by any proof, which can be found only with great effort, a vivid picture is created through simplification and valuation." (Fleck, 1979)

And there is, in these several failures of the received models, methods, and criteria, considerable evidence of a Kuhnian "crisis" in the received praxis of radiation biology: (According to the physicist/philosopher Thomas Kuhn a crisis in science occurs when a severe anomaly is encountered in the practice of Normal science; e.g., when experiments do not produce the results anticipated from the received theory, e.g., "Roentgen's discovery commenced with the recognition that his screen glowed when it should not" (Kuhn, 1970a). This discovery led to a crisis in 19th century science.) In the present report there are several examples of anomaly: i) a cell survival curve for an LQ model is concave up when it "should not" be; ii) the respective dose-response curves for mammary neoplasia in the female rat for neutron and gamma radiation are parallel when they "should not" be; iii) the multifraction LQ model does not fit either experimental or non-experimental data that it "should" fit; iv) the residuals plot of the LQ model of cell survival shows a strong third-order pattern that "should not" be. There are many other examples of current anomalies that are also described in the report. They suggest the presence of a "crisis" in the current praxis of radiobiology. (Or, if "crisis" is thought to be too strong a word, at the least it must be said that there are severe disjunctures in what the case actually is and what it is commonly believed to be, in a disturbing number of instances.)

These "validated reviews" of the 40 studies selected from the radiobiological literature that are presented in this report support several recommendations of measures that will, in the future, assure the quality of the models of radiation dose-response - and isoeffect - that must be deployed in both clinical and public health fields:

a) All published papers must either include the primary data of the study or else the investigators must be obliged to provide it, as requested, at a nominal cost - for which they would be reimbursed.

b) A more self-critical attitude toward their own work and a more skeptical attitude toward that of others must be inculcated in all investigators. The methods of Bayesian estimation and regression diagnostics that are described in the present report (including the Annexes) provide the required insights, perspectives, and criteria, as well as computational techniques to implement such criticism: The techniques of modern statistical modelling, for example, Bayesian regression and regression diagnostics, may be considered to provide a "newer organon" for the investigator.

c) All investigators must take data more seriously. Their conclusions and recommendations must become less theory-driven and more data-driven ("Models are not the source of information, however; data are." R. Thisted, 1980):

i) We must take more data, e.g., more animals at risk at each of more levels of dose, more patients in each "arm" of a clinical trial, etc. Most studies appear to be "data-starved".

ii) We must take more data more often. Data "age", and the substantive conclusions of too many current publications are based upon data that are not only weak, but that were obtained 10-20 - even 30 - years earlier.

iii) We must take more informative data. For one example, we must obtain data on dose-response rather than isoeffect. (N.B.: dose-response models are causal, whereas isoeffect models are deterministic; that is, in the former the predictor variables are measures of the cause of the observed level of radiation effect while in the latter the predictor variables merely determine the level of dose that will elicit a given level of effect. But, "... in the development of a science it is causal relationships which are at a premium." M. Oakes, 1990) For another example, we must obtain information on the joint probability of concomitant occurrence of control of tumor (say, system Q1) and complication of normal tissue in the target volume (system Q2); that is, we must obtain estimates of $\pi_{12}{}^{10} = P(E_1 \text{ and } \bar{E}_2|\underline{x}^T)$ where E_1 represents the binary event ablation of tumor and $\bar{E}_2$ represents the binary event, no complication of normal tissue, and $\underline{x}^T$ is the vector of treatment variables and covariates. (Subscripts denote sites; superscripts identify events.)

iv) We must take more information from that data which is already taken. In Ehrenberg's locution, we must, "Let the data speak" - and then listen attentively. (Ehrenberg, 1975). Moreover, as Welsch (1986) has remarked: "A regression [model] is constructed using prior knowledge, data, models, and a fitting (estimation) process of some form. It is important to know when the resultant regression [model] depends heavily on a small part of the prior knowledge, on a small part of the data, or on the exact choice of model or fitting process."

d) But investigators must not take data too seriously; in the view of a "well-tempered" Bayesian, "Let the data incrementally affect your opinion." (Leamer, 1978)

e) All investigators should deploy the concepts, methods, and criteria of statistical modelling in both the acquisition and the primary analysis of radiation response data.

f) Investigators should be encouraged to become more problem-oriented and less discipline-oriented, since the most insistent - as well as the most interesting - problems rarely respect disciplinary boundaries.

g) Medical physicists must consider that their larger professional role may be - to borrow from Marquardt's (1987) recommendation to statisticians - "purveyors of the scientific method" to the medical profession. In this time of increasing (over?) specialization they must become scientific generalists who can, "practice science - not a particular science" (Bode, Mostellar, Tukey and Winsor, 1949).

The report concludes with an agenda for programs of further work to be undertaken by the Task Group. These are 1) the continued production of "validated reviews" of the literature, including where feasible, a meta-analysis - that is a synthesis, or integration - of either the data or the statistics (e.g., the slopes, the p-values, etc.) of two or more studies of a given issue in dose-response modelling; 2) a Bayesian hierarchical meta-analysis (DuMouchel and Harris, 1983) of animal and human studies on radiation dose-response in order to stabilize (the latter studies "borrow strength" from the former) the estimates of the parameters of the models that are to be deployed in the radiation oncology clinics and radiation protection councils; and 3) In particular, these will be directed to strengthening the parameter estimates of dose-response models of data in which the joint response of two, or more, systems in the irradiated organism is of principal interest to the investigator.

The report has over three hundred pages of text and includes over one hundred and fifty figures. The general format of the report is as follows:

The report includes the main body, sections 1-17, and Appendices I and II and Annexes I-IV. These describe anatomizations of 40 published studies and reviews of studies on radiation dose-response. The Annexes I-IV to the report have been published separately from the rest of the report. Annex I is included in Prediction of Response in Radiation Therapy: Part 1. The Physical and Biological Basis. "Some Impressions of the 3rd ICDTF. A Newer Organon?" 367-377. B.R. Paliwal, J.F. Fowler, D.E. Herbert, T.J. Kinsella, and C.G. Orton, eds. AAPM Symposium Proceedings No. 7. AIP. NY. 1989. Annexes II-IV are included in Prediction of Response in Radiation Therapy: Part 2. Analytical Models and Modelling. "Reflections on the LQ Model. Does it 'Fit'? (Does it Matter?)" 400-517, "Dose-Response Models: Construction, Criticism, Discrimination, Validation and Deployment" 534-630. "Some Applications of Semi-Bayesian Methods to Dose-Response Modelling" 660-717.

The Annexes II-IV provide detailed secondary analyses of 34 of the 40 studies reviewed for this report. These studies describe the received models of both high dose and low dose radiation effects; that is, models of mutagenesis, carcinogenesis, toxicity, and lethality. The results of these analyses are summarized in sections 1-15 of the main body of the report. Section 16 is an epilogue to the report and includes the analyses of three additional related studies published before 1987 on the mouse-to-man problem for high dose effects, which were done after sections 1-15 were completed. Section 17, the coda, describes the secondary analyses of three other additional studies published since 1987. Sections 16 and 17 provide a cross-validation of the findings presented in the earlier sections of the report. Section 16 and 17 present detailed secondary analyses of each of the six studies evaluated therein and includes both tabular and graphical summaries, as do the secondary analyses presented in the Annexes II-IV. We emphasize that the main body of the report, sections 1-15, is only an overview of, and a perspective on, most, though not all, of the material on the 34 studies presented in the Annexes. These latter should be consulted for the detailed results, including tabular summaries, of the respective secondary analyses of each of the studies that the task group evaluated. Appendix I presents the analyses of the classical papers of von Essen on volume effects that provided the initial stimulus for the formation of the task group. Appendix II presents a heuristic discussion of a concept of clinical tolerance that is rather different from current concepts of that phenomenon. The assessments described in sections 16 and 17 serve to reinforce the robust skepticism toward the published radiobiological literature that was developed from our quantitative review of those studies published prior to 1987.

Sections 1-17 of the report are organized around the answers to the four questions on radiation biological models that motivated the work of the task group: What do we believe? Why do we believe it? Should we believe it? What shall we do now? These four questions devolved from the original, more pragmatic - but more intractable - question initially addressed by the task group: What are the nature and size of the losses, say l_{jk}, that may be incurred in clinical decisions based on the assumptions that the received model M_j, with parameter vector β_j, obtains, in the circumstances specified, when, in truth, it is the rival model M_k, with parameter vector β_k, that obtains? The answers to the third and fourth of these questions were sought in the deployment by the Task Group of the concepts, methods, and criteria of modern statistical analysis, especially modern regression analysis (to answer the third question) and meta-analysis (to answer the fourth). The second question has to do with the problems of knowledge and belief. These are the problems of, "... the sources, warrants, and degrees of certainty of scientific findings, the interplay between fact and belief, and between perception and understanding." G. Holton, (1978). Unsurprisingly, a cogent answer to the second question requires some excursions into the philosophy and sociology of science. The answers to all four questions are based upon the material presented in Annexes I-IV and Appendices I and II.

By way of beginning to provide answers to these four questions, the TG1 sought to determine the degree to which the received models were defensible (not merely plausible, since, "... there are different thresholds for the ascription of plausibility") against rival models of the same data set. (As Hippocrates himself remarked some little while ago, "One must attend in medical practice not primarily to plausible theories, but to experience combined with reason.") Or, in other words, to what degree is the received explanation of what the data of a given study mean beyond dispute by reasonable men and women? (Or at least, "... beyond dispute by those who, if not altogether reasonable, are reasonably well-informed and are endowed with reasonable levels of intellectual skills.")

Our examination of the recent and current radiobiological literature disclosed the presence of many severe weaknesses in the studies reviewed. Indeed, those investigators who do not read that literature too closely (and it would seem from our reviews that their number cannot be ignored), may be surprised at just how many things can go (badly) awry in currently published studies on the conception, construction, evaluation, and deployment of the currently received radiation dose-response models. These weaknesses could be roughly, but usefully, identified as ontological weaknesses and epistemological (vide supra) weaknesses (using a classification that is identical to that used for flaws in the analysis and interpretation of biological and medical experiments that was described by Edmond Murphy - a physician at Johns Hopkins - in a 1982 paper. It is, of course, as Yates (1987) has remarked, "... very difficult for scientists to write about philosophy or epistemology so that other scientists will not mis-understand or resent the attempt.") These weaknesses appear to be generic errors of the peer group, rather than the eccentricities of the odd investigator, since similar weaknesses were found in several different studies on quite different endpoints by different investigators working at different institutions at different times.

The ontological weaknesses include the failures of nearly all investigators to recognize and properly account for in their models, either the multivariate or the contingent, or aleatory, aspects of the biological responses of interest. (With respect to the aleatory aspect, the idea that their experiments could have "turned out" differently, and by amounts that can be estimated from their data, does not appear to enjoy much favor.) These failures lead to studies in which the effects of multicollinearity (in the distribution of variables), non-uniformity (in the distribution of observations), and of random (sampling) "noise" in the observed responses combine to grievously degrade the usefulness of the estimates and inferences obtained from their models of radiation dose-response. The epistemological weaknesses include the failure of nearly all investigators to deploy statistically adequate concepts, methods, and criteria in the construction of models, in the assessments of the goodness-of-fit of model and sample and of the consistency of sample and non-sample (a priori) information on the parameter vector, and in the construction of the point and interval estimates of model parameters and functions thereof, such as the response. These failures lead to the deployment of models which are not concordant with the sample data and are inconsistent with non-sample information on the model parameters and functions thereof as well, and to the use of estimates and inferences for which there are insufficient empirical and theoretical warrants.

The correction of these weaknesses in the current radiobiological praxis is not altogether easy since it demands no small degree of familiarity and competence with the concepts, methods, and criteria of matrix algebra and statistical - especially regression - analysis. But it is the case, of course, that good models may be rare because bad models - and the methods by which they are constructed and evaluated - are so easy.

The reader of the TG1 report will find that for many of the studies reviewed, all published in the peer-reviewed literature, our secondary analyses have disclosed that the received beliefs in these fields are in fact confuted by the very evidence that is presented for their support in the published report, when that evidence is assessed in a secondary analysis by a statistically adequate methodology. We cite several of the findings from these analyses as examples:

1) The BEIR III LQ-L model of leukemia incidence (NAS/NRC, 1980), constructed on the LSS sample (T65D dosimetry), overfits the data; there is no sample evidence that supports the addition of the quadratic term, $D\gamma^2$. Examination of both aggregate statistics (such as Pearson chi-squared statistic) and case statistics (such as the plot of the chi-squared residuals vs D_γ) discloses that the rival L-L model is the model of choice (N.B.: It is evident from Table V-8 of the BEIR III report itself (NAS/NRC, 1980) that the LQ-L model overfits these data).

2) The LQ model of cell survival underfits some well-known rat stem-cell survival data; in one case there is strong sample evidence that a cubic term in dose, D^3, must be added to the linear predictor. As in the case of the BEIR III models of leukemia incidence, the evidence resides in the respective Pearson chi-squared statistics and the plots of the chi-squared residuals vs D for the rival models.

3) Since the presence of the redundant quadratic term, D_γ^2, in the LQ-L model of leukemia incidence inflates the variance, Var($\hat{\underline{\beta}}$), of the sample estimate of the model parameter vector, the bias in the sample estimate of the cross-over dose $\theta = \beta_1/\beta_2 = 117$ rad is quite large. The reduced bias estimate (obtained by weighted jackknife methods) is $\hat{\theta}_W = 17$ rad.

4) The so-called α/β ratio for the LQ model, $\alpha/\beta = \beta_1/\beta_2$, is unique only to within a multiplicative constant; therefore it is a poor discriminator of the shape of the dose-response curve of the LQ model of cell survival. We found that the LQ models of cell survival curves with quite similar values of α/β can have quite different "shapes".

5) The dose-response curves for the LQ model of cell-survival are in some cases concave up, i.e., have positive curvature, rather than concave-down (the received shape). Our analyses show that the (received) negative curvature of the LQ cell-survival curves may, in some cases, be an artifact of the semilog plots that are conventionally used to display all cell-survival curves.

6) The parameter vector $\underline{\beta}$ of the LQ model of bone marrow stem cell survival is not invariant between mouse and rat; however, the parameter vector of the rival target theory model is invariant, suggesting that this model is more law-like.

7) We have shown that the conventional methods of estimation of the parameters, say $\alpha(=\beta_1)$ and $\beta(=\beta_2)$, of LQ cell-survival curves that are reported in the literature are equivalent to imposing a constraint, $\beta_0 \equiv \ln(m_1)$, on the parameter estimate, $\hat{\underline{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, of the model, where m_1 is the response at dose, $D_1 = 0$. This constraint, which is equivalent to assigning an infinite weight, $w_1 \longrightarrow \infty$ (say 10^6) on the observation at $D_1 = 0$ is, obviously, difficult to defend. One consequence is that, as well as the point estimates of the model parameters, the interval estimates, or confidence limits, on the estimated response are distorted: At $D_1 = 0$, all confidence intervals for the constrained estimates are zero. But at $D_1 = 0$, the confidence intervals should be near their maximum width - in the absence of the constraint. Such a constraint is, of course, specious, and thus imparts a specious degree of precision to the reported point estimates of the parameters.

8) The conventional Shellabarger et al, 1969 linear, non-threshold, model of radiation-induced mammary neoplasia in Sprague-Dawley female rats is incorrect on the basis of both a priori and

sample information. The linear model implies a uniform tolerance distribution of dose, which is, of course, impossible. We have shown that the sample evidence for the linear model was obtained by "trimming" (by the authors) of the original data; that is, by deletion of the two extreme observations of the original sample, one at dose D = 0 and one at dose D = 500 cGy, the data were forced to fit the linear model.

The correct model of these binary data has a sigmoid dose-response which is, of course, a threshold model both in principle at the D(0.50) = 200 cGy and in practice where the threshold is an effective no-observed effect limit (E-NOEL) at D – 25 cGy.

9) The neutron RBE for radiation-induced mammary neoplasia in Sprague-Dawley female rats is independent of dose, contrary to the received wisdom on this issue, e.g., Montour, 1977. That is, the correct, probit, model gives dose-response curves for this end-point for both neutron and gamma radiation that are parallel.

10) The BEIR III estimates of the parameter vector $\underline{\beta}$ for the LQ-L model of the non-leukemia cancer mortality rates are, in fact, posterior (mixed) estimates of $\underline{\beta}$ for which (we have shown) 40% of the information on dose-response comes from the LSS leukemia incidence data. Since the natural history and radiation dose-response of leukemia (a non-epithelial tumor) are vastly different from those of non-leukemia cancer (an epithelial tumor) the reported (posterior) estimates of the latter are the result of an interspecies transfer of dose-response functions of dubious validity.

11) The "straightness" of the so-called Fe-plot is a very questionable criterion for assessing the goodness-of-fit of the multifraction LQ model of radiation toxicity (hind-leg paresis). We have shown that although in the published study the LQ model is reported to "fit" on this criterion, in fact it does not fit on statistically adequate criteria that are based on the aggregate and case statistics constructed from the Pearson residuals for the binary dose-response model. (We have also shown that in one of the most famous of such studies, the experimental design was such that the F_e-plot would be "straight" regardless of whether the LQ model "fit" the data or not.)

12) The designs of the experiments are badly flawed in most of the studies reviewed: a) The sample data included both outlying and influential observations as well as multicollinear variables. b) The numbers of animals at risk at each level of treatment variables are too small by a factor between 3 and 6. With respect to the first flaw, we have shown that in several studies the estimates of the α/β ratio for the LQ model that are obtained from the F_e-plot are dominated by single observations in the region of $d = D/N \;\underline{-}\; 15$ Gy, well-beyond the stipulated (Fowler 1984, 1989) range of validity ($0 < d < 10$ Gy) of the LQ model. We show that the deletion of this extreme observation changes the sample estimate of α/β by as much as a factor of 2 in each case.

13) The BEIR III estimates of the parameter $\underline{\beta}$ for the LQ-L model and of the cross-over dose $\theta = \beta_1/\beta_2$ are dominated by a single anomalous observation, i.e., deletion of this observation changes the sample estimate of $\theta = 117$ rad to $\theta_{(12)} = 362$ rad, more than a factor of 3.

14) The received estimates for exponents of the so-called tumor-significant dose (TSD) model (0.18, 0.06) are inconsistent with the sample data for which they are reported. We obtain consistent estimates, (0.18, 0.14), for this model by the method of Ridge regression. See Montgomery and Peck, 1982. These estimates lie within the 0.90 confidence limits on the Least Squares estimates and hence are acceptable on the Obenchain (1977) criterion.

15) The LQ model of radiation mutagenesis does not fit the Sparrow et al 1972 data on the induction of pink mutants in Tradescantia 02. These data are heterogeneous and provide an instance of a version of Simpson's paradox: "One of the potential hazards of basing an estimate or test statistic on pooled data from several studies is known as Simpson's paradox. When this paradox arises, the conclusion reached in the meta-analysis may contradict the conclusions of the contributing studies. ... Often the results we want to combine are not homogeneous; in that case ... the biggest problem is that we may be introducing an analogue of Simpson's paradox; when true effects vary from study to study, the size and even the direction of the combined result may depend heavily on such extraneous features as which studies were largest and may hence tend to dominate the analysis" (K. Halvorsen,1986). Our secondary analyses of the Sparrow et al 1972 data disclosed that the observations obtained in their experiment #6, in which the range of dose is 0-12

cGy, dominated the estimates of the parameters of the regression model of the pooled data in which the range of dose is 0-100 cGy.
16) Fowler, 1989 has asserted that, "It should be stressed that no time factor is required for late effects ..." But we have demonstrated that addition of a time factor, log T, significantly improves the goodness-of-fit of both rival models of a well-known set of data on spinal cord injury.
17) The Travis and Tucker, 1987 LQ + time model (the ESD) is not a statistically adequate model of radiation pneumonitis in radiotherapy patients. a) The low dose limb of the 0.95 CL on the lower quantiles of the ESD (that are of clinical interest), e.g., the 0.05 and 0.10 quantiles, are negative, which requires that the appropriate logit model be $z = \beta_0 + \beta_1 \log(ESD)$ rather than $z = \beta_0 + \beta_1 ESD$ as proposed in the Travis and Tucker original report. b) The model of the LQ + time hypothesis that is implied in the 1987 report, $\pi = \alpha D + \beta D^2/N + \gamma T$, where π is the probability of a binary response entails a rectangular distribution of "tolerance dose" which is, a priori, highly implausible. The sample estimates of α, β, and γ obtained from the clinical data yield estimates of $\pi > 1.0$, which is also implausible. c) The sample estimates of the parameters of the logistic model of the clinical data for the LQ + time hypothesis, $z = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3 T$, are not significantly different from zero (and only the estimate of β_3 even exceeds its standard error). This implies that most of the information on clinical dose-response in parameter estimates of the Travis and Tucker model of the clinical data represents that obtained from the animal experiments of Wara et al, 1973 and Field et al, 1976. d) The choice of γT as the form for the time factor is not consistent with the Field et al animal data from which it was estimated (the plots of $D(\alpha/\beta + D/N)$ vs T are "concave down", not "linear" as reported in Travis and Tucker, 1987 when the data of the individual experiments that comprise the pooled data are closely examined - both visually and with a LOWESS "smooth" - separately.
18) The so-called Wara model (Wara et al, 1973) is not a statistically adequate model of radiation pneumonitis in the Mah et al radiotherapy patients. a) As was the case for the ESD of the LQ + time model, the low dose limb of the 0.95 CL on the lower quantiles of the ED (that are of clinical interest), e.g., the 0.05 and 0.10 quantiles, are negative which requires that the appropriate probit model is $z = \beta_0 + \beta_1 \log(ED)$ rather than $z = \beta_0 + \beta_1 ED$ as proposed in the original report (Mah et al, 1987). b) On the basis of the appropriate log(ESD) metameter, the 0.95 CL on the 0.05 quantile are 5.21 Gy in width. There are other problems with the model as well.
19) The LQ + time model does not fit the Withers et al 1988 data on tumor response in head and neck cancer (59 centres). For the Fowler 1991 data on normal tissue radiation response in head and neck cancer (18 centres) the LQ + time model does not "fit" as well as the rival Power-law model. In addition, in the Fowler 1991 study, like the Travis and Tucker 1987 study, the LQ + time model does not satisfy common-sense boundary conditions.

Although it is widely believed that an LQ model can provide an adequate description and explanation for more than one radiation response, our secondary analyses suggest that, in fact, these models can provide an adequate description and explanation for less than one such response, as the above list discloses. We must note that although these, and others of our findings, have disclosed that on the evidence of those (paradigmatic) studies that we reviewed, the models now used to convey information on the relations between radiation dose, time, and response often do not have much in the way of scientific foundations to recommend, or even to defend, them, yet we have not proposed any new models. This position may not, of course, be agreeable to very many of our colleagues. But it must be recalled that the charge to the Task Group did not require that we propose a new model; we were only required to (very carefully) assess the validity of each of the extant models. (Indeed, to require that the weaknesses in the currently received models be ignored until a more adequate rival model can be constructed is not unlike forbidding anyone to point out that the Emperor is, so to speak, buck-naked, unless a full suit of clothes is hanging ready in the closet.) Philosophically speaking, it may be said that TG1 was required to be concerned more with the justification than with the discovery of models. However, given the present nature and level of the empirical knowledge in this field, and the practices by which it is obtained, a disagreeable outlook seems about right. Nonetheless, the report does describe, at some length, the generic defects in the study designs by which current empirical knowledge is acquired and in section 7 and Appendix II describes several rival designs for animal experiments and clinical trials, respectively.

The report is rather long - 350 pages of text and tables and 173 figures - and includes many diverse topics. A brief overview of the thrust and substance of the report can be obtained from the Introduction and summary in section 1 of the report. A brief overview of the statistical concepts, methods, and criteria upon which the arguments, conclusions, and recommendations of the report are founded may be obtained from Statistical Methods I and II presented in sections 4 and 7, respectively. These can be, and indeed have been, used in a graduate-level course in statistical methods for medical physicists. (More precisely, the course is based upon the statistical precepts taken from sections 4, 7, and 9 and the radiobiological examples taken from the set of 40 studies listed in Table 1.) A broad, but still rather detailed appreciation of the entire report can be obtained from a careful study of the Figures, of which there are over one hundred and seventy, most with rather detailed legends. Taken together, the set of Figures comprises a Teaching File for dose-response modelling.

Several of the comments of those members of the Biological Effects Committee (other than the members of TG1) who read the penultimate draft of the report, together with the respective rejoinders, are presented below:

One reader has observed that the statistical concepts, methods, and criteria that were deployed in the secondary analyses of the studies that were presented in this report, and which are described in sections 4 and 7 of the report are "esoteric". (More than one reviewer has suggested that the TG1 must "simplify" this report, but unfortunately few of these recommendations were either very clear on how this should be done or very convincing on either the relative size or the relative value of the simplified report that might be achieved thereby.) It is indeed the case that much of the knowledge in the field of radiobiology that is represented in current published original papers and authoritative reviews can be accurately described as "exoteric", or popular, science, a concept and term attributed to the pathologist/philosopher Ludwig Fleck (Fleck, 1979). This characterization is discussed in section 11 of the report. We have identified and described in the preceding ten sections of the report the characteristic weaknesses of received theory and practice that have required TG1 to classify much of the current knowledge and practices in radiobiological modelling as "exoteric" and have remarked on the lamentable effects of these weaknesses on the received estimates and inferences, which are currently provided in the literature. (The consequences of the clinical deployments of unwarranted estimates and inferences may, of course, be quite grave.)

Another reviewer has remarked that the report is "too long". But the report discloses, study by study, across the entire field of radiation biological effects, what, exactly, is "wrong" (with the evidence, both theoretical and empirical, that is currently presented for the several respective received models, and the estimates and inferences derived therefrom), why it is "wrong", and what is "right" and why it is "right". That is, the report not only identifies the weaknesses in a study and their respective effects on the estimates and inferences presented in the study, but it also shows why these weaknesses - and their effects - are so. The report then presents the more statistically adequate concepts, methods, and criteria and shows why these rival concepts, etc., are the better choices. Although this takes some time, it seemed to us to be the best, if not the only, way to achieve the renovations of the theory and practice that are demanded for the construction and deployment of "useful" models of radiation dose-response. Moreover, it is the case, of course, that any reader needs read only as much of the report as he chooses; but however much he reads, the chances seem quite good that he will know more about dose-response modelling when he puts the report down than he did when he picked it up.

A few of the reviewers were concerned that the findings of the Task Group are "negative". It is the case, of course, that the findings are negative; but there are three positive aspects to our negative findings. One is that only negative findings carry logical implications; that is, only a single negative finding is sufficient to refute an hypothesis, but even a very large number of positive findings will not suffice to prove an hypothesis. (See Popper, 1965a.) Two, although our secondary analyses failed to confirm the published findings of the primary analyses in nearly every study that we reviewed," - failure to confirm an anticipated relationship can represent a contribution of major importance, re-opening questions that had been prematurely closed" (R. Smart, 1964). Several of the anticipated relationships that were not confirmed were: 1) The LQ model of hind leg paresis; 2) The LQ-L model of leukemia incidence (T65D dosimetry); 3) The linear model of mammary neoplasia in the Sprague-Dawley female rat; 4) The LQ model of mutagenesis in Tradescantia; 5) The LQ model of bone marrow stem cell survival; 6) The von Essen model of "volume effects". The third positive aspect is found in a view strongly held by many with respect to any proposed renovation; it can be expressed as the familiar recommendation, "If it's not broken, don't fix it!" From this it would appear that it is necessary to demonstrate persuasively that current radiobiological models and methods are "broken" - or, at the very least, badly "bent" - before many investigators will develop much enthusiasm for the rival, better, methods and models that are presented in the report. This is especially true if, as in the present case, the better methods that are required are also the (much) more difficult and demanding methods. (As remarked above, good models are so rare because bad models are so easy. Or, as John Bailar (1982) - currently a MacArthur Fellow and recently named to the National Academy of Sciences - has observed, "The search for scientific truth is not easy. The questions are often simple. The methods by which we find the answers may not be.") Finally, an observation made in 1978 by the physician, Avedris Donabedian, on the quality of medical care, is quite pertinent to our "negative" findings: "... the only conclusion that can be drawn about levels of quality in general is that whenever the quality of care has been examined, serious and widespread deficiencies have been found" (Donabedian, 1978). Perhaps our findings are not that unusual.

A few other readers have remarked that many physicists and radiation biologists may not read the report since it is "too statistical", as well as both "long" and "esoteric". It is, of course, a truism that many radiation biologists and physicists - as well as many other scientists - do not often choose to read statistical monographs - as indeed many do only rarely read statistical textbooks or statistical journals, either. The current view of statistical methodology that is widely held by nearly all biological and physical scientists is nicely captured by two lines from W. H. Auden: "Thou shalt not sit with statisticians/Nor commit a social science." That this is surely the case is, as remarked repeatedly in the report, evident to even a casual review of the recent and current radiobiological literature which, as we have repeatedly pointed out, is obviously much the weaker, or "exoteric", for these volitional defects. Indeed, if that were not the case, then the findings and conclusions of the TG1 report would likely have been altogether different. But, and not incidentally, it is quite evident from the findings of the TG1 report that a surprising number of investigators may not really read the scientific literature either, even those studies which they may cite - and seek to emulate. Or, if they are actually reading that literature, it is apparently with small result, since, as remarked above, their conclusions are often severely weakened, or even in some cases, completely confuted, by the evidence in the studies that are cited.

However that may be, it seems to us that the assumption that the majority of physicists and radiation biologists, upon learning of the egregious consequences of the weaknesses in current radiobiological praxis will not rouse themselves to address the matter by learning the better methods - and why they are better - is as unfair as it is uncertain. It may be that these conjectures - by several reviewers of the TG1 report - that the report will enjoy only a small readership are correct ("You can lead a horse to water ..."?). However, we believe that the radiobiology and radiation physics communities deserve the benefit of the doubt on this issue.

Moreover, it is the case that although this task group report may well require that the investigator devote no little time to its careful study, it could also prove to be a most economical use of that time since it will reduce the numbers of papers in the radiobiological literature that he will find necessary to read thoroughly. For unless a paper exhibits those statistical - and logical - warrants for the "believability" of its reported results that are identified, instanced, and recommended in the present report, it is not only unnecessary, but even unwise, for the investigator to read the paper too closely - except perhaps for whatever interest it may hold for him as a cultural artefact. Indeed, this feature may well be one of the chief values of the TG1 report. For as the mathematician David Hilbert once remarked, the importance of a scientific paper can be judged by the number of previous publications it makes unnecessary to read. One measure of the value of the TG1 report is the number of both previous and subsequent publications it makes unnecessary to read. The state of the current literature suggests that of the relentlessly increasing number of papers that are now published there are very few that will be found to reward careful study. Thus, on the basis of the "Hilbert measure" the TG1 report appeals directly to that behavioral extremum principle known as the Principle of Least Effort, or Zipf's Law, after the Harvard philologist, George Kingsley Zipf (Zipf, 1965). Hence, it should prove to be a valuable addition to any serious investigator's professional library.

We have sought to motivate the required changes in the methodology that is currently deployed in the construction, etc., of radiobiological models by demonstrating that failure of investigators to learn and to deploy the appropriate, "esoteric", statistical concepts, methods, criteria, and standards, is a practice that is not only unscientific ("Now it is possible to maintain ... that statistics in its broadest sense is the matrix of all experimental science and is consequently a branch of scientific method, if not Scientific Method itself; and, hence, that it transcends the application of the scientific method in sundry fields of specialization. The scientist should know statistics as he knows logic and formal language for communicating his ideas" Kendall, 1968. Or, to paraphrase Lord Rutherford's views on Physics: "All Empirical Science is either Statistics or it's stamp-collecting."), but that it is unethical, as well ("So what is the relation between statistics and medical ethics?" ... "Stated simply, it is unethical to carry out bad scientific experiments. Statistical methods are one aspect of this. However praiseworthy a study may be from other points of view, if the statistical aspects are substandard then the research will be unethical." Altman, 1982). N.B.: In order to be evenhanded in this matter, we must observe that although the sweet reasonableness of the statistical insights and perspectives, as well as of the statistical concepts, methods, and criteria that we describe - and proffer - in the report seems to the members of the TG1 to be altogether obvious, this is, apparently, not the case for a significant group of significant scientists. Their views are nicely captured by a comment from Claude Bernard, the eminent 19th century French physician who is often deferred to as the father of modern experimental medicine: "But, it is not with the aid of statistics that we shall reach it [scientific realism]; statistics never has and never could tell us anything about the nature of phenomena." C. Bernard, 1865. However, we must also remark that since modern statistics began only in the second decade of the present century - and modern regression analysis (including Bayesian regression) did not begin to develop until the 1960s, it may well be the case that Dr. Claude and Sir Maurice (Kendall) are both right. Certainly, Dr. Bernard's view's were better informed (although incorrect) in his century than are the views of those who still adhere to them.

By way of further motivation we have also shown that the relentless accumulation of "exoteric knowledge" in the professional radiobiological literature may well lead to an embarrassment of the entire profession in some future courtroom in which the presiding judge requires instead, esoteric knowledge, from the expert witness who is testifying.

A few readers have objected to the (admittedly) unconventional format followed in the TG1 report. Although, the report format was determined largely by the requirements of the charge to the task group, it was also, to a large degree, determined by our previous experience with more conventional formats, such as that followed in the monograph "Multiple Regression Analysis: Applications in the Health Sciences", D. Herbert and R. Myers, editors, AAPM monograph No. 13, American Institute of Physics, New York, NY 10017, 1986, which have, apparently, not been as successful in inducing the required sea-change in the statistical practices employed by many of those investigators concerned with the construction, testing, and deployment of radiobiological models as was intended. Therefore, we felt compelled to offer an alternative format that would stimulate and provoke, as well as instruct and inform. Indeed, as Richard Dawkins has remarked (in the "Blind Watchmaker," a modern defense of Darwin's theory of evolution), "... it sometimes isn't enough to lay evidence before the reader in a dispassionate way. You have to become an advocate and use the tricks of the advocate's trade. This book is not a dispassionate scientific treatise ... it seeks to inform, but it also seeks to persuade ..." Just so the TG1 report.

Some recent history suggests that such an alternative format, "... one that seeks to inform but ... also seeks to persuade ...," is now warranted - or even overdue: Since the mid-nineteen sixties there have appeared in the biomedical literature the results of surveys of that literature which document the occurrence of appalling levels of statistical malpractice therein. We quote from just three of these studies as examples:

1) Schor et al (1966) in their review of ten medical journals remarked that, "Among the 149 analytical studies critically evaluated ... less than 28% were considered acceptable ... Five percent were judged unsalvageable: the problem posed by the investigator could not be solved by the kind of study described. A majority should have been revised before publication. Thus, in almost 73% of the reports read (those needing revision and those which should have been rejected), conclusions were drawn when the justification for these conclusions was invalid."

2) Glantz (1980) remarked that, "approximately half the articles published in medical journals that use statistical methods use them incorrectly. These errors are so widespread that the present system of peer review has not been able to control them."

3) Williamson et al (1986), in their review of 28 reviews of the quality of the medical literature (such as Schor et al, 1966), found that, "... the 12 [reviews] published during or after 1970 reported that a median of 6 percent of 2172 publications met the assessors criteria;" "... the assessments usually involved leading medical journals, such as the New England Journal of Medicine, the Journal of the American Medical Association, and the Lancet. ... Since the leading journals tend to be more selective than others, the quality of the entire medical literature is likely to be worse than these findings indicate." ... "Overall, these findings may be typical for three reasons. First, poor scientific quality was found in nearly all the research - regardless of type and content - encompassed in these 28 assessment articles. Second, when design, analysis, and documentation were assessed concurrently, the proportion of articles meeting the criteria often dropped to less than 1 per cent. Third, similar findings have been confirmed in the physical science literature. In our opinion, with respect to the clinical literature in English, the findings of the 28 assessments probably overstate the scientific quality of research publications in the applied health sciences."

A more recent comment from the dean of the UCSD School of Medicine suggests that there has been no abrupt change in the levels of statistical malpractice that encumber the biomedical literature since the Williamson et al, 1986 paper: "The quality of statistical analysis (something of a disgrace in the literature) could be improved, ..." P. J. Friedman, 1988.

As any reader of the TG1 report will soon discover, the findings of the TG1 on the quality of the literature on radiation dose-response models presented in our report are consistent with those of Schor et al (1966), Glantz (1980), and Williamson et al (1986) (And, in a few cases, with those of Friedman (1988) as well.)

Such findings raise an insistent question that must concern all of us; namely, what are the underlying causes of the continued egregious performance in a field - cancer research and treatment - which has never lacked for practical and economic motivation and support. (For example, the NCI has expended well over $22,000,000,000 over the past two decades - i.e., the period covered in the reviews of Schor et al, Glantz, and Williamson et al - in fighting the"War on Cancer".)

It would appear that the underlying cause is not a dearth of statistical insights, perspectives, and appropriate methodologies: Since the late nineteen-fifties there has been an explosive increase in the size and sophistication of the statistical armamentarium available to researchers through textbooks, journal articles, etc., particularly in regression modelling - which surveys have shown to be the most widely used of the statistical methodologies available. ("As a rough guess, about 10^5 data sets per day are used as input to multiple regression programs around the world." A. Miller, 1983. And, "The developments in regression methodology in the past 25 years lead one to wonder about the quality of the analyses performed prior to this recent quarter-century [that is, 1959-1982]." Hocking, 1983. Rather detailed expositions of several of the most important of these "developments in regression methodology in the past 25 years," can be found in the AAPM monograph 13 cited above.)

The repeated findings of "statistical malpractice" in the published reports of biomedical studies over a twenty-five year period that includes an explosive development in statistical methods and insights suggest the possibility that the criticisms presented in the biomedical literature and correctives provided in the statistical literature are both being ignored by many biomedical researchers. On the other hand, it is more likely that the researchers are either unaware of the poor statistical quality of their work and the consequent weaknesses in their inferences and estimates or that, although perhaps aware of it, they do not really have a realizing sense of the problem and how to correct it because there is as yet no single convenient document available to both motivate and inform them. Such a document is provided by the TG1 report - including its Annexes. That is, as remarked above, the TG1 report provides detailed and vividly illustrated expositions of what is wrong (with much of current radiobiology praxis), and why it is wrong, and what are the correct statistical concepts, methods, and criteria, and why these are correct. The examples of incorrect concepts, etc., serve to motivate the study of the correct concepts, etc., that can inform subsequent studies by the investigator. (At least one or two of the published studies evaluated in the TG1 report are not simple incorrect, but, are, in fact, absurd; to borrow a phrase from Sir Ronald Fisher, these studies include "conspicuous and catastrophic howlers" and hence have great motivational value: They show that it is possible to publish and perish - if one's report should be read by someone with a sharp pencil.)

A shorter version of the TG1 report would devolve into simply a "naming of (defective) parts." It would merely be another documentation and inventory of instances of statistical malpractice in an already lengthy jeremiad which began with Schor et al in 1966. The complete report by TG1 provides some remedy for - rather than just another description of - the problem of statistical malpractice, that is, it presents an exposition, together with an abundance of practical illustrations, of statistical concepts, methods, and criteria for assuring the quality of radiation oncology practice in an area of that practice in which physicists (e.g., Goitein, Hall, Orton, Schultheiss) have long performed with distinction: The construction and deployment of regression models of radiation dose-response data.

Perhaps a very few (two or three?) investigators will acknowledge the reported findings that, in general, the "statistical analyses are something of a disgrace" - and even unethical as well (see section 10 "Statistical methods and the ethics of scientific enquiry") - and still think "So what?" since, of course, a great deal has been learned about the biological effects of ionizing radiation and, moreover, by the usual "bottom-line" criteria of scientific success - grants, journals, promotions, publications, tenures, etc. ("riches, honour, and the love of women") - the radiation biology enterprise has been rather successful. If, however, the mortality rates of the common cancers, surely an important criterion, are used as a broader index of the success of the enterprise, a rather different conclusion obtrudes. For example, S. Dische, in his Mackenzie Davidson Memorial Lecture, "Advances in basic science: have they benefited patients with cancer?", published in the British Journal of Radiology, noted that in the UK "... deaths due to neoplasms as a percentage of all deaths have risen from 20.2% in 1970 to 25.1% in 1986" (Dische, 1991b) - an increase by a factor of 1.24. In this lecture he observes that one of the more robust and professionally fructifying preoccupations of radiation biology for well over fifty years (beginning with J. C. Mottram in 1935) has been the elucidation and exploitation of the so-called "oxygen effect." To this latter end especially, over 70 randomized clinical trials have been directed (Dische, 1991a, b). Nevertheless, "We must acknowledge that so far we have failed to advance the care of the cancer patient" (Dische, 1991b).

More generally, J. Bailar and E. Smith observed in their NEJM 1986 paper, "Progress Against Cancer?" (quoted in Annex I to the TG1 report) that, "The main conclusion we draw is that some 35 years of intense effort focused largely on improving treatment must be judged a qualified failure. Results have not been what they were intended and expected to be." Bailar and Smith's conclusions, rancourously challenged (by some) when first published, now appear to have been corroborated by a GAO study reported in a 1992 issue of Science: "... overall death rates from many common cancers remain stubbornly unchanged - or even higher than when the war [on cancer] began." In this context it is of relevance to note that recently the JCAHO has begun to change their criteria for the credentialing of health organizations from "structure and process" to "outcome". Thus, the issue has become not can a health organization - or enterprise - deliver the desired result, say "quality care", but rather, does an organization or enterprise deliver the desired result (US Congress, 1988). JCAHO credentialing now demands evidence of actual performance (at an acceptable level) such as the organizations' (risk-adjusted) mortality rates, as in the HCFA reports since 1988 (Medicare Hospital, 1992), or, the Pennsylvania Health Council CABG report of 1992 (Pennsylvania Health, 1992), rather than evidence of the organizations' capacity to perform (at an acceptable level). As Arnold Relman, editor of the New England Journal of Medicine has noted,"We can no longer afford to provide health care without knowing more about its successes and failures. The Era of Assessment and Accountability is dawning at last; ..." Relman, 1988. See also Elwood, 1988.

In their 1986 paper Bailar and Smith remark that, "we think that there could be much current value in a comprehensive, consolidated, objective review of the technical reasons for this failure ... Why were hopes so high, what went wrong, and can future efforts be built on more realistic expectations. Why is cancer the only major cause of death for which age-adjusted mortality rates are still increasing?" To the degree that models of radiation dose-response motivate and inform the management of cancer by radiation treatment the TG1 report may help to provide a "... comprehensive, consolidated objective review of [some of] the technical reasons for this failure." To the same degree the TG1 report also suggests, "Why hopes were so high," and, "What went wrong?" The TG1 report also describes throughout (and especially in section 14), "What shall we do now?") some "future efforts" that are built on "more realistic expectations."

Finally, several reviewers have remarked on the abundance of quotations that appear throughout the report. (At least one internal reviewer recommended that all quotations be deleted. Two others suggested that they should be collated and published separately.) We have indeed deployed a large number of appropriately selected quotations - or epigrams, or aphorisms, or apercus - as a rhetorical device to introduce, to summarize, to motivate, to give salience to, and to justify, or even to legitimate, the positions taken in each of the several sections of the report. The use of such a device seems warranted given that the topics presented may appear to be difficult, the positions taken unfamiliar and often in opposition to those that comprise the received wisdom, and the conclusions disagreeable to perhaps a large number of readers. But, of course, such a procedure can be regarded as simply a way of "... following the common scientific practice of the persuasive use of citations to construct [our] case" (S. Kellert, 1993). And the information presented in these quotations may be more accessible - it is certainly more concise - and the positions taken and conclusions reached found more acceptable if couched in, or buttressed by, the comments of acknowledged authority-figures. For some sections we have felt it necessary to muster a Greek chorus of such commentators both to lay down the story-line and to justify the denouement.

Donald Herbert, Ph.D.

Quality Assessment and Improvement for Dose-Response Models.
Some Effects of Study Weaknesses on Study Findings.
"C'est Magnifique?"

Table of Contents

N.B. Annex I is published in Prediction of Response in Radiation Therapy: The Physical and Biological Basis. Part 1. B.R. Paliwal, J.F. Fowler, D.E. Herbert, T.J. Kinsella, and C.G. Orton, eds. Proceedings of the 3rd Intl. Conf. on Dose, Time and Fractionation in Radiation Oncology. Sept. 14-17, 1988. Madison, WI. AAPM Symposium No. 7. AIP. NY. 1989.

Annexes II-IV are published in Prediction of Response in Radiation Therapy: Analytical Models and Modelling. Part 2. B.R. Paliwal, J.F. Fowler, D.E. Herbert, T.J. Kinsella, and C.G. Orton, eds. Proceedings of the 3rd Intl. Conf. on Dose, Time and Fractionation in Radiation Oncology. Sept. 14-17, 1988. Madison, WI. AAPM Symposium No. 7. AIP. NY. 1989.

Quality Assessment and Improvement for Dose-Response Models. Some Effects of Study Weaknesses on Study Findings. "C'est Magnifique ?"

A Report of the Biological Effects Committee Task Group 1,
"Evaluation of Models for Dose-Response in Radiation Oncology."

Donald E.Herbert, Ph.D., principal author and task group chairman 1984-1992

Arnold Feldman, Ph.D.
Engokolai Krishnan, M.D.
Colin Orton, Ph.D.
Jacques Ovadia, Ph.D.
Bhudatt Paliwal, Ph.D.
Timothy Schultheiss, Ph.D.
Prakash Shrivastava, Ph.D.
Alfred Smith, Ph.D.
Marilyn Stovall, MPH
Lionel Cohen, M.D., Ph.D. (Consultant)

"'There remains something in all of us of the childish belief that there is a world of grown-ups who know. There must be - because we evidently don't know.' [P. Berger, 1961] Herein lies the explanation for the historically authenticated role of experts: they are the 'grown-ups who know'."

I. Hoos, 1980

"It can scarcely be questioned that when the truth or falsehood of an event or observation may have important bearings on conduct, over-doubt is more socially valuable than over-credulity ... One of the most fatal (and not so impossible) futures for science would be the institution of a scientific hierarchy which would brand as heretical all doubt as to its conclusions, all criticism of its results."

K. Pearson, 1892/1957

"Indeed, history shows that skepticism is preferable in science."

K. Rothman, 1986

"I know that most men - not only those considered clever, but even those who are clever and capable of understanding the most difficult scientific, mathematical, or philosophic problems - can seldom discern even the simplest and most obvious truth if it be such as obliges them to admit the falsity of conclusions they have formed, perhaps with much difficulty - conclusions of which they are proud, which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives."

L. Tolstoy, 1898

"The trick is not so much to know something, but to know you know it."

I. Stewart, 1990

"The central problem of epistemology always has been, and still is, the problem of the growth of knowledge ... [Science] is an alternating series of speculative conjectures and empirical refutations."

K. Popper, 1965b

1. Introduction and Summary

Recently, there have been several evaluations, by various academic, judicial, and legislative bodies of diverse backgrounds, agendas, and constituencies, of the provenance and probative value of the intellectual product currently retailed in the biomedical literature. See, for example, the (notorious) paper by Stewart and Feder (1987) and the reports of the AAAS/ABA workshops on Project on Scientific Fraud and Misconduct (AAAS/ABA 1988, 1989a, 1989b), as well as the several articles and editorials in NEJM, Science, JAMA, Lancet, Cell, Radiology, et cetera, concerning the Baltimore, Darzee, and Slutsky affairs. The results of these evaluations strongly suggest that more than the usual prudence may be the most appropriate policy for any consumer of that literature. For example, Williamson et al (1986) remark that, "The findings ... suggest that the average practitioner will find relatively few journal articles that are scientifically sound in terms of reporting usable data and providing even moderately strong support for their inferences." Moreover, "The mere fact that research reports are published, even in the most prestigious journals, is no guarantee of their quality." And, in a published discussion on randomized controlled clinical trials (RCTs), Collins (1987) has noted that, "... we shouldn't believe the results published - even in peer-reviewed journals. ... you need to check the data that are published in much more detail." In the same publication, Chalmers (1987) remarks that, "We've published many analyses showing that randomized controlled clinical trials as published in peer-reviewed journals are often of very poor quality."

Some aspects of still another basis for counseling prudence are examined in the present report from Task Group 1 which provides a summary of the findings of a recent and extensive critical review of some of the radiation oncology/biology literature. Specifically, the report addresses the consequential issue of the degree to which the received[1] estimates and inferences that have been made from some of the received models of radiation dose-response and are reported in that literature, should be accepted as "believable"; that is, "capable of instilling faith, trust, or acceptance (a believable explanation)". The criticisms presented are logical as well as methodological in character. The main thrust of our criticisms is not so much directed to refute certain of the received models of dose-response - although this is certainly one consequence (vide infra) - but rather to show that certain concepts, methods, and criteria of modelling are inadmissible - methods and criteria which have been widely used without being challenged for many years in the construction, interpretation, and deployment of dose-response models in radiation biology and oncology. The reader will find that a battery of arguments has been ranged against that prime manifestation of radiobiological orthodoxy: the linear-quadratic model. These arguments point out (among other things) that in the service of that orthodoxy an astonishing number of transmogrifications of data and methods and criteria and standards have been pressed.

The preceding drawing by Gary Larson is, of course, a caricature of this situation, but like all caricatures it is a distortion that is developed from a grain of truth: The TG1 report discloses that many of both current and previous findings reported in the radiobiological literature, that is, the estimates and inferences concerning the several responses of biological systems to ionizing radiation, were and are obtained by a concatenation of "wrongs", that is, by the systematic application of dubious methods to questionable data. We present some cogent arguments for a sea-change in the perception and evaluation of familiar data and, more generally, in the processes by which knowledge is currently acquired, evaluated, and deployed in this field. The report may be said to be pre-occupied with the fundamental and insistent epistemological question of how we can learn from data - or fail to.

We note also that many (though not all) of these beliefs - anent the received concepts, methods, and criteria of radiobiology - seem to have become "cultural truisms"; that is, beliefs that are so widely and unquestioningly held that their adherents (a) are unlikely ever to have heard them being questioned and may therefore (b) have difficulty defending them against such criticism (Greenwald, 1975).

While polemic is surely no stranger to academe, a disclaimer must be entered at this point. Although the present report identifies and comments on several weaknesses in the concepts, methods, criteria, and standards that inform and animate the current practices of radiobiological science, as disclosed to our review of selections from the radiobiological literature, no personal criticism of individual scientists who may be, or wish to be, identified with these concepts, etc., is implied - or intended. But constructive criticism, to be effective, must be aimed at what is defective rather than at what is laudable. And, indeed, "One principle of scientific scholarship is to learn from the misadventures and follies of those who preceded us on the investigative scene" (A. Feinstein, 1990). However, our arguments are ad opus, not ad hominem. The report is meant as a challenge to thinking - not to personal combat. The TG1 seeks to stand on the shoulders -

not on the faces - of those distinguished scientists whose published work is reviewed herein. We initially intended to prepare a quantitative and integrative review, that is, a meta-analysis, of such work. However, this intent was frustrated by the weaknesses we found in most ofthe studies that we reviewed.

1.1 Meta-analysis

"Meta-analysis is a new discipline which critically reviews and statistically combines previous research."

T. Chalmers, 1987

"Meta-analysis ... spotlights the defects of the original research it seeks to combine." "... there are grave defects in the majority of original research that only come to light if an attempt is made to combine the data with that of other research."

T. Chalmers, 1989

"[Meta-analysis] is going to revolutionize how the sciences, especially medicine, handle data. And it's going to be the way many arguments will be ended."

T. Chalmers, 1990

"Scientific research is founded on integration and replication of results; with the possible exception of a new discovery, a single study rarely makes dramatic contributions to the advancement of knowledge."

S. Thacker, M.D., 1988

"... the combination of individual estimates is not a routine matter, but requires clear thinking about both the nature of the data and the function of a combined estimate."

W. Cochran, 1954

"This surprising phenomenon, that one can combine unrelated problems to get improved estimators, is what is commonly called the Stein effect."

J. Berger, 1983

"Simply put, the Stein effect asserts that estimates can be improved by using information from all co-ordinates when estimating each co-ordinate."

G. Casella, 1985

"Meta-analysis is ... an attempt to rigourously evaluate what we have learned (if anything at all), to uncover what we must do to learn more, and to provide mechanisms to 'force' the necessary improvement."

K. O'Rourke and A. Detsky, 1989

"A single clinical trial often fails to give clear-cut generalizable results, because of insufficient patient numbers and the particular way the study was done. Meta-analysis (or overviews) use formal statistical techniques to combine the results from similar trials not only to increase numbers but also to generalize conclusions to a more varied range of patients and treatment protocols."

S. Thompson and S. Pocock, 1991

Our evaluation of the literature of radiation biology may be usefully viewed as the first critical step in a recently developed general procedure in the "study of studies" known now as meta-analysis. The latter is a set of quantitative methods for summarizing and integrating the results (either data or statistics) from several - or even many - related studies of a common hypothesis to arrive at conclusions, e.g., rejection of the hypothesis, that could not be achieved otherwise because of various weaknesses - such as small sample sizes - in the component studies. Meta-analysis is a new scientific discipline that critically assesses and subsequently combines the results of several previous studies of a given issue. Classical meta-analysis was introduced into medicine as a post-hoc salvage maneuver for defective multicenter randomized clinical trials. However, most studies - animal experiments as well as clinical trials - that are reported in the medical literature today are data-starved. For example, the median number of patients, n_i, assigned to each arm of the typical randomized controlled clinical trial is about 20, whereas the required number of patients is 100-200 (Pocock and Hughes, 1990; Zelen, 1982). Similarly, the number, n_i, of animals (typically small rodents) at risk at each level of exposure in a radiation dose-response experiment in which the binary end-point is not mutagenesis or carcinogenesis (for instance it may be a radiation toxicity such as paresis) is typically about 5 whereas 30-50 animals are required - preferably in 3 or more replicates of 10 or more animals each (Finney, 1971b). (These findings suggest that many investigators do not take data very seriously - since they take so little of it.)

"The purposes of meta-analysis include the following: (1) to increase statistical power for primary end-points and for subgroups, (2) to resolve uncertainty when reports disagree, (3) to improve estimates of effect size, and (4) to answer questions not posed at the start of individual trials" H. Sacks et al (1987). It is an important "salvage maneuver" because there are a) so many clinical trials that are b) undersized. These have the following effects: a) As the number of trials of an ineffective treatment increases the probability of obtaining a false positive result increases (the "multiple comparisons problem"). b) Under-sized trials have low statistical power to detect clinically significant differences between rival treatments increasing the probability of obtaining a false negative result. (Oakes 1990 shows that if the statistical power (i.e., $1 - \beta$) of a study is low, say 0.50, then the percentage of false positives for that study can greatly exceed the designed level (which is typically $\alpha = 0.05$). In Oakes' example, for $n = 5000$, $\alpha = 0.05$ and an achieved power of 0.50, the percentage of false positives may be 30%.) Meta-analysis can decrease the probability of both types of errors, although there are, to be sure, some weaknesses of the method, e.g., heterogeneity of the studies combined and/or publication bias. The latter is present whenever the studies actually published in the peer-reviewed journals comprise a small and biased sample of the studies actually completed on a given issue (See section 2.2 below). Meta-analysis of clinical trials can also improve the validity of the extrapolation and generalization of the results of the trials (i.e., the so-called Phase IV clinical trial).

"In most methodological and applied papers in the biomedical and epidemiologic literature, meta-analysis is undertaken to achieve one of two aims: Obtaining a pooled estimate of an overall 'true treatment effect' with associated standard error and confidence intervals or 'borrowing strength' from related studies to sharpen the estimate of an effect of interest in a particular study or class of studies." (NAS/NRC, 1992). It will be useful at this point to briefly comment on the issue of simultaneous estimation and the phenomenon of the so-called Stein effect since it relates to the possibility of studies "borrowing strength" from one another. In 1955 C. Stein established the paradoxical effect that the usual estimator of a k-variate Normal mean, namely, the vector of sample means, say $\underline{x}^T = (x_1, ..., x_k)$, was inadmissable, that is, it could be improved upon, for $k \geq 3$ (Stein, 1956). To put it simply, the Stein effect states that the classical maximum likelihood sample estimates, $\hat{\theta}_i$, of k multivariate parameters, θ_i, $1 \leq i \leq k \geq 3$, can be improved by using information on all k parameters when estimating each parameter. The improvement is measured in terms of squared loss, $\Sigma(\theta_i - \hat{\theta}_i)^2$. This loss may be represented as a sum of a variance and a bias term. The Stein effect achieves a reduction in mean-squared error by reduction of the variance at the expense of an increase in bias, i.e., it is a reduced-variance estimator. The Stein effect has been found to hold in essentially any normal estimation problem involving the estimation of at least a three-dimensional mean. ("In fact, Stein-type effects probably occur in most multiparameter estimation problems ..." B. Efron and C. Morris, 1973a.) Independence of co-ordinates is not needed, and hence the results apply to quite general estimation in the normal linear model: for instance, the ordinary least squares estimator of three or more regression coefficients in a linear regression can be improved upon via the stein effect. The James-Stein estimator performs well in prediction because it anticipates regression toward the mean.

It is of course the case, that in order to take practical advantage of the Stein effect, some prior information about the θ_i must be taken into account. But, although it is closely related to Bayesian ideas in most applications, it is important to realize that the Stein effect is a distinct phenomenon. This is so because the θ_i need not be related in any way, or even be the same type of parameter, for the Stein effect to hold. However, the related empirical Bayes or hierarchical Bayes, estimators to be discussed below in section 14.3, do require the θ_i to be related in some way and these latter techniques can readily estimate the degree of relationship - the "novel error of uncertain relevance" from the data itself (DuMouchel and Harris, 1983).

In the more common version of meta-analytic inquiry the investigators use information on the features of the individual studies to account for the variation in the respective outcomes of the several studies. That is, meta-analysis is used to estimate both the within-study variability and the between-study variability.

In the TG1 report we sought to identify those study features - study contexts, designs, subjects, methods and criteria - that account for the unique outcome of the individual study; in particular we seek to identify - and quantify - the effects of study weaknesses on study findings. To anticipate some of our results we must note that we found that in several cases the study outcome could be explained as simply a methodological artifact, a finding that tends to cast doubt upon the published interpretations of what the data really mean in many of the studies examined. Therefore, although the major goal of meta-analysis is to statistically combine the results of previous research into a more defensible conclusion (or estimate, or inference, or model) we found early on in our study that the published results on the issues of interest in the studies that we examined were of too poor a quality to be meaningfully integrated - by any means. Therefore, we report only the critical reviews of individual studies that must, of course, precede any integration or synthesis of results. (As DuMouchel (1990a, b) has noted, "... many published studies do not contain enough statistical detail to enable the parameter estimates, and, especially, their standard errors, to be calculated. A particular author may have taken a different analysis tack than that required by the proposed meta-analysis, or even misanalyzed the data. To construct the needed set of parallel estimates with associated standard errors it may be necessary to reanalyze many of the original data sets." An attempt at a meaningful synthesis will be made when, or if, stronger primary studies of the matters at issue can be found. (See part 14, "What shall we do now?"). However, several of the studies that were included in our critical review provide examples of quite simple meta-analyses that achieved varying degrees of success. (None of them, however, were originally reported as meta-analyses.) One example is the pooling of the city-specific data on the Hiroshima and Nagasaki survivors in the LSS sample of the BEIR III (1980) report. Another example is the pooling of the data from four different experiments on mutagenesis that is reported by Sparrow et al (1972). A third example is the pooling by Travis and Tucker (1987) of the data of the five experiments on mouse pneumonitis by Field et al (1976) and Wara et al (1973). (It is the case, of course, that these two studies are more precisely described as instances of data augmentation rather than meta-analysis. However, both procedures have the common aim of enabling weak studies to "borrow strength" from one another - or from stronger studies.)

The evaluation of the literature that was undertaken, as initially conceived and described in the (revised) charge to the Task Group 1 of 1985, was restricted in scope to the critical evaluation of those models of dose-response currently deployed in clinical radiation oncology. However, it became apparent almost at once that both the insights and perspectives on the identity and effects of study weaknesses on study findings that were to be achieved by such a critique would be greatly enhanced by a critical evaluation of the received modelling praxis across the whole spectrum of radiation dose-response studies: mutagenesis, carcinogenesis, toxicity, and lethality.

There are a number of arguments for the value of such a panoramic view from which between-study similarities and differences may be remarked, most of which will be obvious to the thoughtful reader. Therefore, we will cite only the most cogent: 1) when the respective responses are considered according to the inherent nature of the conditional distribution of the random part of the observed response it is found that for all studies this is either a Poisson or a Binomial distribution. Therefore, the respective dose-response models have either common or similar forms. For example, a) the models that implement the LQ hypotheses of mutagenesis, or carcinogenesis, and lethality are Poisson linear and Poisson log-linear, respectively; b) the multivariate probit model is appropriate to convey the information on dose-response that is contained in the data of either toxicity or neoplastic transformation studies in which the observed response is a proportion and hence the distribution of response is Binomial. 2) Although studies of the quantal response at high doses - radiation toxicity - and at low doses - radiation carcinogenesis and mutagenesis - are both animated by a concern to model the probability of occurrence of the respective binary events, the reported received models of the former are isoeffect models and those of the latter are dose-response models.

Therefore, critical evaluation of the models of two sets of seemingly disparate responses enhances the understanding of the received concepts, methods, and criteria deployed in both sets of studies. It permits us to examine common issues of modelling praxis independently of issues of endpoints of response. In particular, it assists in the discrimination between idiosyncracies that are peculiar to the practices of an individual study and more general weaknesses in the paradigm that informs and guides the research program of the whole field.

This larger view has disclosed that the presence of what may be termed a Kuhnian crisis in the received praxis of radiobiology modelling is more real than was apparent from the initial and more restricted study of models of dose-response in radiation oncology. It will be recalled that the historian/philosopher Thomas Kuhn identifies as a crisis in science those rare circumstances in which the paradigm that has hitherto informed and guided the practice of so-called "Normal Science" by the peer-group is no longer adequate to these purposes. That is, the existing scientific tradition embodied in the received paradigm can no longer provide the required "puzzles" that are to be solved by the peer-group in their practice of Normal science since a new experience has disclosed the presence of an anomaly, that is, a failure of anticipation, a severe mis-match between the actual occurrences of events - either in level or in kind - and their expectations as based on the received paradigm. ("Under normal conditions the research scientist is not an innovator but a solver of puzzles, and the puzzles upon which he concentrates are just those which he believes can be both stated and solved within the existing scientific tradition" ... "Bringing a normal research problem to a conclusion is achieving the anticipated in a new way, and it requires the solution of all sorts of complex, instrumental, conceptual, and mathematical puzzles. The man who succeeds proves himself an expert puzzle-solver ... Though intrinsic value is no criterion for a puzzle, the assured existence of a solution is." T. Kuhn, 1970)

1.2 Validated reviews

The results of these evaluations of the respective empirical foundations for the received opinions on the radiation dose-response relations in mutagenesis, carcino-leukemo-genesis, toxicity, and lethality, as disclosed in over three dozen peer-reviewed and frequently cited studies and authoritative reviews of such studies, have been published in AAPM Symposium Proceedings #7 (1989). See Table 1. These analyses are described as Annexes I-IV of this report. This set of analyses may be usefully viewed as yet another effort at providing the "validated review" of the literature that has been recommended by Williamson et al, 1986, in the review cited above, as one remedy for the poor quality, remarked above, of both the published research reports, and authoritative reviews of those reports, that comprise that literature: "..., to keep up with a given field, seek reviews based on validated research sources. Such articles describe the scope of the literature surveyed and indicate the criteria applied to determine the scientific adequacy of the source information presented the review." (emphasis added). These re-evaluations of the individual studies have come to be known as secondary analyses. In this more recent context we must then add that, "Primary analysis is the original analysis of data ... It is what one typically imagines as the application of statistical methods. Secondary analysis is the re-analysis of data for the purpose of answering the original ... question with better statistical techniques or answering new questions with old data." The "... benefits [of secondary analysis] include the verification and refinement of original findings and the refutation of them." (It is most important to note that meta-analyses of a set of studies must usually be based upon the results of preceding secondary analyses of each of the component studies, since in so many studies the primary analyses will be found to be flawed. In this regard, it should be remarked that for this reason any given meta-analysis may not

Table 1. Published Papers and Reviews Evaluated by TG1 (The number of citations to 30 June 1993 are given in parentheses).

"... we gain our knowledge of science by reading and thinking about other people's experiments ..."

J. Ziman, 1968

"... there is no substitute for a careful, or even meticulous, examination of all original papers purporting to establish new facts."

R. A. Fisher, 1936

1. Alper, T. Survival curve Models. IN Radiation Biology in Cancer Research. 3-18. R.E. Myers and H.R. Withers, eds. Raven Press. NY. 1980. (10)
2. Ang, K.K., van der Kogel, A.J., Dam, J.V. and van der Schueren, E. The Kinetics of Repair of Sublethal Damage in the Rat Cervical Spinal Cord During Fractionated Irradiations. Radiotherapy and Oncol. 1: 247-253. 1984. (37)
3. Bentzen, S.M., Christensen, J.J., Overgaard, J. and Overgaard, M. Some Methodological Problems in Estimating Radiobiological Parameters from Clinical Data. Acta Oncol. 27: Fasc. 2. 105-116. 1987. (9)
4. Bond, V.P., Cronkite, E.P., Lippincott, S.W. and Shellabarger, C.J. Studies on Radiation-Induced Mammary Gland Neoplasia in the Rat. Radia. Res. 12: 276-285. 1960. (42)
5. Chapman, J.D. Biophysical Models of Mammalian Cell Inactivation by Radiation. IN Radia. Biol. in Cancer Research. 21-32. R.E. Meyn and H.R. Withers, eds., Raven Press, NY. 1980. (16)
6. Cohen, L. The Tissue Volume Factor in Radiation Oncology. Int. J. Radia. Oncol. Biol. Phys. 8: 1771-1774. 1982. (15)
7. Comas, F.V. The Radiosensitivity of Rat Bone-Marrow Cells. Int. J. Radia. Biol. 17(6): 549-557. 1970. (11)
8. Fertil, B., Deschavanne, P.J., Lachet, B. and Malaise, E.P. In Vitro Radio-Sensitiv Human Cell Lines. Radia. Res. 82: 297-309. 1980. (40)
9. Fertil, B. and Malaise, E.P. Inherent Cellular Radiosensitivity as a Basic Concept for Human Tumor Radiotherapy. Intl. J. Radia. Oncol. Biol. Phys. 7: 621-629. 1981. (97)
10. Fertil, B. and Malaise, E.P. Intrinsic Radiosensitivity of Human Cell Lines is Correlated with Radioresponsiveness of Human Tumors: Analysis of 101 Published Survival Curves. Intl. J. Radia. Oncol. Biol. Phys. 11: 1699-1707. 1985. (140)
11. Field, S.B., Hornsey, S. and Kutsutani, Y. Effects of Fractionated Irradiation on Mouse Lung and a Phenomenon of Slow Repair. Br.J. Radiol. 49: 700-707. 1976. (118)
12. Fowler, J.F. A Critical Look at Empirical Formulae in Fractionated Radiotherapy. Biological Bases and Clinical Implications of Tumor Radioresistance. 201-204. G. Fletcher, et al, eds. Masson Pub. NY. 1983. (4)
13. Fowler, J.F. What Next in Fractionated Radiotherapy? Br. J. Cancer. 49: Suppl. VI. 285-300. 1984. (76)
14. Fowler, J.F. Fractionated Radiation Therapy after Strandqvist. Acta Radiol. 23. Fasc. 4. 209-216. 1984. (18)
15. Fowler, J.F. The Linear-quadratic Formula and Progress in Fractionated Radiotherapy. Br. J. Radiol. 62: 679-694. 1989. (35)
16. Fowler, J.F. Apparent Rates of Proliferation of Acutely Responding Normal Tissues During Radiotherapy of Head and Neck Cancer. Int. J. Radia. Oncol. Biol. Phys. 21: 1451-1456. 1991. (1)
17. Frome, E.L. and Beauchamp, J.J. Maximum Likelihood Estimation of Survival Curve Parameters. Biometrics. 24: 595-605. 1968. (8)

18. Keane, T.J., Fyles, A., O'Sullivan, B., Barton, M., Maki, E. and Simm, J. The Effect of Treatment Duration on Local Control of Squamous Carcinoma of the Tonsil and Carcinoma of the Cervix. Seminars of Radia. Oncol. 2: 26-28. 1992. (0)

19. Mah, K., Van Dyk, J., Keane, T. and Poon, P.Y. Acute Radiation-Induced Pulmonary Damage: A Clinical Study on the Response to Fractionated Radiation Thearpy. Int. J. Radia. Oncol Biol. Phys. 13: 179-188. 1987. (18)

20. Millar, B.C., Fielden, E.M. and Millar, J.L. Interpretation of Survival-Curve Data for Chinese Hamster Cells, Line V-79 Using the Multi-target, Multi-target with Initial Slope, and α, β Equations. Intl. J. Radia. Biol. 33(6): 599-603. 1978. (45)

21. Montour, J.L., Hard, R.C. and Flora, R.E. Mammary Neoplasia in the Rat Following High-Energy Neutron Irradiation. Cancer Research. 37: 2619-2623. 1977. (8)

22. NAS/NRC The Effects on Populations of Exposure to Low Levels of Ionizing Radiation: 1980. (BEIR III) National Academy Press. Washington, DC. 1980. (62)

23. NCRP Influence of Dose and Its Distribution in Time on Dose-Response Relationships for Low-Let Radiation. NCRP Report No. 64. Washington, DC. 1980. (78)

24. NIH. Report of the National Institutes of Health Ad Hoc Working Group to Develop Radioepidemiological Tables. NIH Pub. No. 85-2748. US Gov't. Printing Office. Washington, DC. 1985. (0)

25. Shellabarger, C.J., Bond, V.P., Cronkite, E.P. and Aponte, G.E. Relationship of Dose of Total-Body 60Co Radiation to Incidence of Mammary Neoplasia in Female Rats. IN Radiation Induced Cancer. 161-172. IAEA. Vienna. 1969. (19)

26. Sparrow, A.H., Underbrink, A.G. and Rossi, H.H. Mutations Induced in Tradescantia by Small Doses of X-Rays and Neutrons: Analysis of Dose-Response Curves. Science. 176: 916-918. 1972. (106)

27. Supe, S.J., Nagalaxmi, K.V. and Meenaksi, L. Tumor Significant Dose. Med. Phys. 10(1): 51-56. 1983. (0)

28. Thames, H.D., Withers, H.R., Peters, L.J. and Fletcher, G.H. Changes in Early and Late Radiation Responses with Altered Dose Fractionation: Implications for Dose-Survival Relationships. Intl. J. Radia. Oncol. Biol. Phys. 8: 219-226. 1982. (236)

29. Till, J.E. and McCulloch, E.A. A Direct Measurement of the Radiation sensitivity of Normal Mouse bone Marrow Cells. Radia. Res. 14: 213-222. 1961. (2828)

30. Travis, E.L. and Tucker, S.L. Isoeffect Models and Fractionated Radiation Therapy. Intl. J. Radia. Oncol. Biol. Phys. 13: 283-287. 1987. (24)

31. Tucker, S.L. Tests for the Fit of the Linear-Quadratic Model to Radiation Isoeffect Data. Intl. J. Radia. Oncol. Biol. Phys. 10: 1933-1939. 1984. (38)

32. Tucker, S.L. and Thames, H.D. Flexure Dose: The Low-Dose Limit of Effective Fractionation. Intl. J. Radia. Oncol. Biol. Phys. 9: 1373-1383. 1983. (28)

33. UNSCEAR. Genetic and Somatic Effects of Ionizing Radiation. 1986 Report to the General Assembly, with Annexes. United Nations. NY. 1986. (8)

34. van der Kogel, A.J. Late Effects of Radiation on the Spinal Cord. Doctoral Thesis. University of Amsterdam. Radiobiological Inst. of the Organ for Health Research. The Netherlands. 1979. (64)

35. von Essen, C.F. Roentgentherapy of Skin and Lip Carcinoma: Factors Influencing Success and Failure. Amer. J. Roentgen. 83: 556-570. 1960. (42)

36. von Essen, C.F. A Spatial Model of Time-Dose Area Relationships in Radiation Therapy. Radiology. 81: 881-883. 1963. (24)

37. von Essen, C.F. Clinical Radiation Tolerance of the Skin and Upper Aero-digestive Tract. Front. Radia. Ther. Onc. 6: 148-159. Karger, Basel and Univ. Park Press. Baltimore. 1972. (2)

38. Wara, W.M., Phillips, T.L., Margolis, L.W. and Smith, V. Radiation Pneumonitis: A New Approach to the Derivation of Time-Dose Factors. Cancer. 32: 547-552. 1973. (175)

39. Withers, H.R., Taylor, J.M.G. and Maciejewski, B. Treatment Volume and Tissue Tolerance. Int. J. Radia. Oncol. Biol. Phys. 14: 751-759. 1988. (31)

40. Withers, H.R., Taylor, J.M.G. and Maciejewski, B. The Hazard of Accelerated Tumor Clonogen Repopulation During Radiotherapy. ACTA Oncol. 27: Fasc. 2. 131-146. 1988. (115)

Notes

1) Of the several papers listed in Table 1, the reports by Till and McCulloch (1961) and Frome and Beauchamp (1968) represent the very best in scientific enquiry and reporting. We have also used the data on which these two studies were based to compare the received concepts, methods, and criteria of radiation biology with those of statistical modelling in the construction of models of dose-response of radiation lethality (cell-survival) data. See Annex II, part 5 and Annex IV, part 6.

Many of the remaining studies provide instructive "concrete examples" (Pearson, 1892/1957) of the "deft resolution of the wrong problem" or of one or more of the following weaknesses:
a) Lack of knowledge of subject matter.
b) Faulty, misleading, or imprecise interpretation of the data and results.
c) Incorrect or flawed basic data.
d) Incorrect or inadequate methodology.
As Jaffe and Spirer (1987) have remarked, "The presence of any one of these flaws is enough to invalidate the results [of any study] ..."

2) Whether any given study is influential in forming and guiding the subsequent research in a given field, cannot, of course, usually be established until several years, say 3-5, have elapsed following its initial publication.

3) All of the primary studies listed above have been cited in at least one authoritative review published in 1985 or later.

4) It has hitherto been accepted that, following publication, a scientific paper is, on average, cited ~ 0.7 times/year. More recently, statistics compiled by the Philadelphia-based Institute for Scientific Information (ISI) indicate that 55% of the papers published between 1981 and 1985 in journals indexed by the institute received no citations at all in the 5 years after they were published. ... ISI's data-base covers only the top science and social science journals - some 4500 out of nearly 74,000 scientific titles listed in Bowker/Ulrich's data base a commercial listing of all periodicals ... An earlier ISI study of articles in the hard sciences (including medicine and engineering) published between 1969 and 1981 revealed that only 42% received more than one citation ... Moreover, self-citation - a practice in which authors cite their own earlier work - accounts for between 5% and 20% of all citation ..." D. Hamilton, Science, 1990. Therefore, nearly all of the papers listed in the table are, by any measure, "well above average". Some are, obviously, classics in their field.

5) In this study we found it more fruitful to anatomize a few selected studies in order not only to identify the obvious weaknesses but also to demonstrate the causes of these weaknesses, their respective effects on the estimates and inferences made from each study, and to suggest maneuvers whereby new studies might be strengthened and old studies salvaged. This feature distinguishes our report from many other contemporary reviews of the medical literature which have usefully identified, inventoried, and reported the often appalling statistical weaknesses in large sections of the medical literature.

Each of the several studies that are listed in Table 1 provide examples of several different aspects of the received theory and practice of radiobiological science and therefore each paper has been cross-classified under several rubrics and examined in several different contexts and from several different perspectives in different sections of this report (as well as in the Annexes). For example, the classic study by Shellabarger et al (1969) on the incidence of radiation-induced mammary neoplasia in the female rat has been examined in the context of low-dose studies, in the context of invariance of estimates of slope-parameters between studies, in the context of the shape of dose-response curves, etc. There is, then, some unavoidable redundancy between different sections of the text as well as, of course, between the main body of the report and the Annexes which it summarizes.

We have also exploited the evidence to be found, in several of the reports listed in Table 1 of the often egregious failures of the received methods and models, to motivate the study of the statistically adequate methods and criteria that are presented in sections 4 and 7. In this endeavor, the reader will note that extensive use was made of the models and methods of the BEIR III (1980) report. It is, of course, the case that as a source of inferences and estimates on the response of populations to low doses of ionizing radiation BEIR III has been superseded by the BEIR IV (1988) and BEIR V (1990) reports, although comparisons of the risk estimates of BEIR III with those of the latter reports still hold some interest for many investigators. However, the BEIR III report will remain a fruitful source of Cautionary Tales for the data analyst for many years. More than the (T65D) dosimetry was found to be dubious in the BEIR III report. Even if the (T65D) dosimetry were correct the estimates and inferences published in the BEIR III report are questionable because of weaknesses in the analysis; e.g., the BEIR III models of choice do not fit the BEIR III data all that well (on the standards and criteria of modern regression analysis, e.g., Robins and Greenland, 1986).

Most of our secondary analyses of the studies listed in Table 1 have disclosed that these radiobiology studies do not provide a defensible scientific foundation for radiation protection policy, nor do they provide much of a scientific foundation for radiation oncology practices - nor, indeed, do they provide much of a foundation for radiation biology itself. And to those who would dismiss our set of examples as "merely anecdotal", we respond that a sufficiency of such anecdotes does describe the a pattern that supports our conclusions.

The following figures were re-drawn (with permission) from their respective sources: Figure 2a, Figure 9, Figure 10a, Figure 22, Figure 30a, Figure 31b, Figure 31c, Figure 39a, Figure 39b, and Figure 45a.
Appendix I: Figure 1 and Figure 2.

The following figures are reproduced with permission from AAPM/AIP publications:

Figure 34b, 35a, 35b, 35c, and 35d (Taken from Herbert, 1981). Figure 1a, 1b, 5a, 5b, 6a, 6b, 6c, 7a, 8a, 8b, 11a, 13a, 13b, 21a, 21b, 21c, 23a, 23b, 23c, 24d, 24e, 24f, 25a, 30b, 30c, 31a, 32, 33a, 33b, 34a, 37, 38a, 38b, and 38c (Taken from Herbert, 1989c). Figure 3, 11b, 11c, 12a, 12b, 14a, 14b, 15a, 15b, 16a, 16b, 16c, 17, 18b, 18d, 18e, 18f, 19a, 19b, 20a, 20b, 20c, 20d, 25b, 26a, 26b, 27a, 27b, 27c, 28a, 28b, 28c, 29, and 34c (Taken from Herbert, 1989b). Figure 2b, 36a and 36b (Taken from Herbert, 1989d). Figure 40a, 41a, 51b, 42a, 42b, 45d, 45e, 45f, 46a, 46b, 47, 48a, 48b, 49, 50a, 50b, 51, 52a,52b, 53, 54a,54b, 55a,55b, 56a,56b, 57a, and 57b (Taken from Herbert, 1993a).

be successful in either providing a more useful estimate or a more defensible decision. Moreover, there have been several well-founded criticisms of meta-analytic philosophy and procedures. However, the results of these secondary analyses that must precede any meta-analysis, whether they confirm or confute the results of the primary analyses, must be regarded, in and of themselves as positive achievements.)

Note also that, in general scientific practice, the confirmation - or validation - of previous findings requires new data in which to test the degree to which these findings can be generalized (See section 9 of this Report). However, contradictions of previous findings are most persuasive when based on the data of the original study. Indeed, it seems to be quite essential that contradictions of previous findings be based on the data of the original study for, as Mazur (1973) has remarked, "A common way to deal with data which are inconsistent with one's own position is to deny their statistical validity." Received data have the essential feature of a demonstrated acceptability by the peer-group as "the facts" even though they may be shown to be highly questionable in one or more other respects. (For, as Leamer (1983) has remarked, "What is a fact? A fact is merely an opinion held by all, or at least held by a set of people you regard to be a close approximation to all" - that is, by the peer-group).

This inherent feature of secondary analysis has helped to set much of the mildly iconoclastic "tone" of the main body of the present report (and of the findings of Annexes I-IV which it summarizes): "Secondary analysis may reanalyze data by following the original researcher's methods, thus checking the accuracy of the reported results, or by using competing analytic techniques or sets of assumptions, thus testing the robustness of the original conclusions to alternative approaches" (as in a sensitivity analysis). Moreover, "If independent reanalyses are done conscientiously and with visibility, the credibility [and thus the importance] of the original research may be enhanced."

It follows that a prominent feature of this work in discriminating between those conclusions to a published study that are defensible and those that are merely plausible, resides in the identification of study weakness and the evaluation of the effects of these weaknesses on study findings: weaknesses in the prior assumptions that motivated and informed the collection of the data, weaknesses in the data, weaknesses in the data analysis, etc. (Unfortunately, there are, of course, "individual thresholds for the ascription of plausibility".) Thus, this review is much more of a quantitative assessment, rather than the more familiar narrative account, of current studies. This quality of (mildly) confrontational skepticism sharply distinguishes this review from recent and current authoritative reviews of the radiation biology/oncology literature which appear to be unduly authoritarian and subjective and which, increasingly, seem to be more a celebration than a critique of the work that is cited. (However, as the philosopher/physicist Thomas Kuhn has remarked, confrontation of opposing views is the only method by which science has ever progressed.) As noted above, Williamson et al (1986) recommend that medical practitioners, "to keep up with a given field, seek reviews based on validated research sources", and explicitly warn them to, "... beware of any review that merely summarizes and does not validate its source information."

The current practice of the authoritative reviews of the radiobiological literature (to celebrate rather than to validate its source information) exacerbates the effects of the failure of the sources cited to point out any weaknesses that may exist in the evidence which they offer in support of their conclusions. But, as the Nobel laureate Richard Feynman has observed, "Details that could throw doubt on your interpretation must be given if you know them. You must do the best you can - if you know anything at all wrong, or possibly wrong - to explain it." However, as Ed Gehan (1983) points out, "When writing up the results of some studies, a delicate problem arises in the emphasis to be given to reporting known 'inaccuracies in the data'. Giving too much detail about a problem can supply critics with a basis for arguing that the results must be necessarily invalid. Not divulging problems would be a cover-up. The only ethical approach is to report problems and inaccuracies, and to evaluate the effect that these might have had on the major conclusions of the study."

Both precedent and exemplar for the type of critical assessment that we offer in this report are provided by the work of the so-called Particle Data Group (Rosenfeld, 1975), first organized in 1957, which publishes (in alternate years) a systematic review of all the research done worldwide on the properties of elementary particles. (It should be noted that this group has found it necessary to delete as many as 40 percent of the published experiments from their overviews because they believe the data are "unreliable, preliminary, incorporate assumptions" that they consider questionable or, yield estimates that deviate too greatly from those of other studies. Such a finding is rather jarring given that both historians (e.g., Thomas Kuhn and Alexandre Koyre) and philosophers (e.g., Rudolph Carnap and Sir Karl Popper) alike have characterized "physics" as the exemplar of Western empirical science. Moreover, as both Kuhn and Popper have pointed out, that's all there is - there is no other kind of empirical science. The discovery of deterministic chaos (Moon, 1992) and the recognition of its implications for the predictability of the behaviour of simple (few degrees of freedom) non-linear physical systems has prompted some recent revisionist thoughts on the future role of physics as the paradigm of Western science by A. Pippard, FRS: "... it is desirable that those social scientists who seem to strive to make their discipline conform as closely as possible to the perceived ideal of physics, should recognize that they may be imitating physical procedures at the very point where they are least reliable. And ... physicists should not allow themselves to be gratified over much when philosophers of science select physics as the typical science. The phenomenon of chaos is a salutary reminder of the failure of human endeavor, and it may be that the recognition of the limitations of mathematical prediction will prove to be the most typically scientific aspect of physics."

Another precedent for the type of critical analysis that we have conducted - and propose to continue - is provided by the work of the Thermophysical Properties Research Center (Touloukian, 1975) that was also organized in 1957 in response to a similar dilemma arising in still another field of applied science; namely, that, in their words, "We do not know what we know." - and thus on this important criterion, the knowledge in their field was weak.

Moreover, such secondary analyses recommend themselves not only by their eminently practical achievements in the validation of fundamental tenets of any applied science but also because they provide an occasion to exercise that essential function - and perquisite - of University teaching: the criticism of commonly received beliefs. In the latter context it has been remarked that, "The study of fallacy in concrete examples ought to play a greater part in our educational curriculum. Certain works have a permanent value in this respect." (K. Pearson, "Grammar of Science", 1892/1957).

1.3 Quality assurance in modelling

The problems of "quality assurance" for dose-response models are essentially the problems of knowledge: "... the sources, warrants, and degrees of certainty of scientific findings, the interplay between fact and belief and between perception and understanding." (G. Holton, 1978). In particular, they are the problems in the processes by which we learn from data. On this issue the message of this AAPM report is simple: The methodology and mind-set required for the quality assurance that can deliver a "quality" manufactured product - a product that meets the needs at the market-place - can likewise be deployed in the service of delivering a quality intellectual product - in the present case, say, dose-response models, that meet the needs of the radiation oncologist and the radiation epidemiologist. The evaluations described in Annexes I-IV were performed as part of an ongoing quality assurance study that was begun in 1984 by Task Group 1 of the Biological Effects Committee of AAPM (well before the studies of Williamson, et al (1986) referred to above were published). It is the case, of course, that both the methods and the numbers that implement and describe, respectively, the effects of the machines (the linacs, the computers, etc.) that are routinely deployed in radiation oncology have long been the object of a systematic and commendably severe scrutiny - and to good effect since events both untoward and actionable can be shown to have occurred occasionally in the absence of such scrutiny (quality assurance is surely an act of the reasonable and prudent - as well as the competitive - man). But curiously enough the models of radiation dose-response by which radiation treatments are motivated and, presumably, informed, have not hitherto been exposed to similar concentrations of "critical scrutiny". In particular one cannot find in the published studies that describe - and often advocate - the deployment of these models, any evidence of systematic and well-informed efforts to identify and assess the effects of possible study weaknesses on study findings. The work described in this report is among the first of such assessments.

We note that although the studies that were evaluated for the present report did not include any randomized clinical trials (RCTs), it was evident to us early on that the findings of the recent evaluations of these trials by other investigators such as Berlin, Chalmers, Simes, Zelen, etc., are quite similar in several respects to our findings in those studies which we assessed. Therefore, we have included summaries of their findings on the RCTs to provide an additional perspective from which to view our own.

These evaluations of the radiobiological literature disclose something of what the radiation biology/oncology community currently sees as legitimate problems and of what it counts as acceptable solutions thereof. ("Scientific journal articles have status not only as reports of results but as statements of how those results were achieved" P. Woolf, 1988.) That is, our evaluations permit us to discern something of the current paradigms (Kuhn, 1970a) that motivate and inform both the thinking and the practices of this group. A paradigm defines the problems available for scientific scrutiny, the standards by which the profession determines what should count as an admissible problem or as a legitimate problem solution (Kuhn, 1970a). The LQ model provides an instance of a current paradigm and the clinical section of the radiobiology peer-group currently sees the obtaining of point estimates of the α/β ratio for a multifraction LQ model of a given tissue as a legitimate - even insistent - problem, and it counts the F_e-plot as an acceptable solution thereto (Fowler, 1984, 1989). However, it is important to note that, "... paradigms are not to be entirely equated with theories. Most fundamentally they are accepted concrete examples of scientific achievement, actual problem solutions which scientists study with care and upon which they model their own work" (Kuhn, 1977).

As remarked above, the secondary analyses that were undertaken by the task group have disclosed the presence of several severe operational and philosophical shortcomings in the current modelling praxis of the radiation biology/oncology group. By way of disclaimer, and without intending to seem overly profound, (As F. Eugene Yates (1987) has remarked, "It proves very difficult for scientists to write about philosophy or epistemology so that other scientists will not mis-understand or resent the attempt.") it must be said at once that the evaluations described in Annexes I-IV were much more concerned with the meta-physical or ontological issues (what - and how much of it - this part of the natural world does and does not include) rather than with the deontological issues that have been addressed by the several so-called "fraud-busters" groups that were referred to at the beginning of this report. The ontological weaknesses that were disclosed include both overstatements and understatements of "what's really out there"; that is, they reside either in the absence of empirical evidence for phenomena (entities, categories or processes) that ought to be there, or, in the presence of empirical evidence for phenomena that ought not to be there, on the basis of currently received opinions. These are, for example, the perdurable questions of a) the putative "shape" - including the presence/absence of "thresholds" and "saturation effects" - of radiation dose-response curves and surfaces, including, of course, the roles of fractionation and protraction in modulating the biological effects of radiation dose, the empirical validity - and theoretical interpretation - of the LQ model as well as the meaning and value of its characteristic parametric function, the so-called α/β ratio; b) the putative dose-dependence of the neutron RBE; c) the existence of a putative "volume effect", etc.

However, the general epistemological issues of how - and how well - do we know what we know (about radiation dose-response) were, perforce, addressed as well in these secondary analyses. Indeed, it was such latter issues that provided the initial motivation to undertake the secondary analyses described in Annexes I-IV of the present report, most of which are described herein. It will soon become obvious to the reader that the TG1 report is pre-occupied with the central epistemilogical issues of, "How we learn from data - or fail to."

1.4 The charge to the task group

The initial concern that motivated the formation of the Task Group in 1984 may be briefly characterized as follows. There are now, of course, several alternative models of radiation dose-response currently deployed in the clinic. For each of these models neither the theoretical nor the empirical evidence is unequivocal. However, some of these models are currently more highly esteemed than are others; these are the so-called received models. An insistent question then arises: What are the nature and size of the losses that may be incurred, say l_{jk}, in clinical decisions based on the assumption that the received model M_j, with parameter vector $\underline{\beta}_j$, obtains, in the circumstances specified, when, in truth, it is the rival model M_k, with parameter vector $\underline{\beta}_k$, that obtains? It soon became clear that cogent answers to this question would require considerable preliminary information; indeed, it now appears that it would require more information -and more analysis - than could be reasonably expected to be accumulated during the usual lifetime of an AAPM task group. Therefore, it was re-framed as four existential questions anent dose-response relations in cells and tissues exposed to ionizing radiation for which answers might be more readily obtained: 1) What do we believe? 2) Why do we believe it? 3) Should we believe it? (Is it "believable"?) 4) What shall we do now? (It may be noted that these insistent queries are similar to three posed earlier by the German physicist/philosopher Kant: What must be the case? What should we do? For what may we hope? (See Hacking, 1983 and Kant, 1800.)

"What we believe" includes a brief inventory of not only the received radiobiological ontology - the hypotheses and "laws", together with the models that articulate and implement them, and the estimates and inferences derived therefrom - but also of the received methods, criteria, and standards by which such knowledge is acquired and evaluated and by which the received models are constructed, evaluated and deployed. It also includes the received inferences and estimates of model parameters, and of the linear and nonlinear functions thereof, such as the predicted response and the α/β ratio, respectively.

"Why we believe" in the received hypotheses, "laws", methods, and criteria is answered by a careful evaluation and interpretation of the evidence presented in the published (peer-reviewed) literature. We have found that often the received evidence is simply either the authority of common practice, that is, practices that are common to, and in some cases even peculiar to, the radiobiology group, or the authority of a single, perhaps charismatic, investigator or institution. On the other hand, many of the published beliefs in the received models, methods, etc., can apparently only be justified on what we have termed "para-normative" (a nonce word) criteria since the empirical evidence offered in many studies is logically devastating to the respective conclusions reported in these studies. We present below some discussion of early and recent work on the formation and maintenance of beliefs by para-normative criteria that seems to explain the difficulty.

The question of whether we should believe the received hypotheses, "laws", estimates, methods, and criteria (i.e., are they "believable"?) is answered by assessing the probative value of the evidentiary basis for them that is presented in the literature and, in particular, by comparing the relative weights of the evidence offered by common and accepted sets of data for received and rival models of that data.

The answers to the fourth question, "What should we do now?", are based upon the findings of our evaluations of the evidentiary basis of the received beliefs and the normative responses to them. These suggest that there are three entirely complementary courses of action that might be profitably followed and we discuss these at some length in part 14 of this report. We note that one course of action, the less-precedented, comprises, in fact, a novel research program that addresses two insistent problems: 1) the interspecies transfer of dose-response models (a Bayesian model of the "mouse-to-man" problem) and 2) the estimation of the joint probability of occurrence of complication in normal tissues and control of tumor in the target volume (a bivariate probit model of the "joint effects" problem).

However, rather than providing specific answers to each of these four questions, as in, say, a listing of the more salient features of the received ontology, together with received methods and criteria by which the received evidence for them has been demonstrated, a listing of the cognate features of the rival ontology together with rival methods and criteria by which the received evidence is refuted - or confirmed -, etc., all of which can be retrieved by the reader from the

Annexes I-IV of this report, we believe that it is more useful to use each of these questions as discussion points around which to organize one or more of the important parts of the exposition of the investigation described in this report.

Thus, the present report, including Appendices I and II and Annexes I-IV, presents brief discussions of a) some preliminary answers to these four questions, together with our evidence for these answers; b) some discussion of the methods and criteria by which these answers were sought, obtained, and validated; c) a brief but important discussion of the relation of the deployment of appropriate statistical methods and criteria to the ethics of scientific enquiry; d) some remarks on the philosophy of scientific inquiry by way of providing something of the scientific and philosophical contexts in which these answers can be best assimilated, understood, and interpreted, as part of a larger and coherent whole - or so it seemed to us. With respect to d), it is, of course the case, that most scientists, even the more distinguished, usually have little interest in either the history or the philosophy of science. ("... most of them are no more concerned with the philosophy of science than most lawyers are with the philosophy of law" A. Woodcock and M. Davis, 1978. Indeed, "... many scientists find an explicit consideration of such matters irritating ... Nevertheless, it's well to recall the remarks of the mathematician David Hawkins: 'Philosophy may be ignored but not escaped; and those who ignore most escape least.'" See J. Casti, 1989.) However, the insights provided by both of these disciplines are important to the questions addressed in this report since the answers depend very much upon an operational definition of science, and a theory of how and why it works. In particular, the contexts provided by the history and philosophy of science help to explain why we found what we did - and not something different - in our secondary analyses of the studies listed in Table 1. Thus, it is of more than passing interest to note that critical comparisons of those writings of the early philosophers of science, the reformers and innovators of the Middle Ages and the Renaissance, in which the maxims, methods, and precepts for the successful conduct of Western science were first laid out, with our findings on current research reports, discloses the presence of an astonishing (and often amusing) level of intellectual recidivism in some reports. That is, several of the habits of thought, methods of argument, and standards of proof that were stigmatized as foolish, or as folly (or worse), in the thirteenth century by Roger Bacon, in the fourteenth century by William of Ockham, in the seventeenth century by Francis Bacon, etc., can be readily identified in the reports by several distinguished scientists that have been published in the latter half of the twentieth century. It was the physicist Maupertuis who defined mechanical inertia as, "the presence of the past"; the current literature provides several examples of a kind of mental interia. Citations from the writings of these early reformers are intended to persuade the reader that our current concerns are not only justified but indeed have an ancient and honorable scientific pedigree.

Fairly recently, Prigogine (1980) has noted that there is now "... a strong current both in Europe and in the United States to bring the philosophical and the scientific themes closer together." Moreover, it is important to remark two areas of professional interest to medical scientists in which there is currently active discussion of the philosophical issues that arise in the practice of science: 1) epidemiology (beginning with Buck, 1975: "Epidemiologists are exceptionally concerned about their method of approach. In few other medical sciences is so much attention devoted to the <u>philosophical</u>, as opposed to the purely technical aspect of method. The reason for this is that in epidemiology the experiment plays a relatively minor role. ... In Popper's terms, ... the data must be capable of <u>refuting</u> the hypothesis ...") 2) the law ("... the problems science poses for the law parallel and reflect the philosophical problem of defining science." Black, 1986). Both are concerned with such seemingly recherché issues as causality, "Hume's problem" (induction), "Kant's problem" (demarcation), the logic of scientific proof - and disproof, etc. In addition, there appears to be a convergence of the respective core issues in the recent developments in the philosophy of science, artificial intelligence, and the analysis of clinical diagnosis (Schaffner, 1985. See also the 1982 paper by G. R. Dolby.)

In the present context of data acquisition and analysis, it must be remembered that the central problem in each instance is always, "What to think after the analysis is complete." (Leamer, 1978; Dempster, 1983). The philosophy - and history - of science provides several important perspectives on any substantive answer to this problem: It is crucial to the understanding of either how or what - and the meaning thereof - we can learn from data. For example, and speaking quite broadly, does the result of the analysis of a set of data represent a refutation - or corroboration - of a conjecture per Popper? Or, per Kuhn, is the result to be interpreted as the solution to one of the puzzles provided by the current paradigm - or as the identification of an anomaly, a novelty of fact or theory, that provides the counter-instance which provokes the crisis that leads to the overthrow of the current paradigm - and its replacement by another in which the anomaly is no longer so? See K. Popper Logic of Scientific Discovery (1965a) and Conjectures and Refutations (1965b) and T. Kuhn The Structure of Scientific Revolutions (1970a) and The Essential Tension (1977). We remark that, as unfamiliar as they may be to many - and as uncongenial as many others may find them - the views of both Popper and Kuhn on the way in which science is practiced in the last half of the twentieth century are quite fruitful - not to say provocative. For example, the reader may ask himself how well the multifraction LQ model, constructed upon an unspecifiable level, S, of the survival of an unidentifiable - and indefinable - (single) population of cells satisfies the Popper criterion, "falsifiability" ("In so far as a scientific statement speaks about reality, it must be falsifiable: and in so far as it is not falsifiable it does not speak about reality."), that demarcates science (e.g., Einstein's relativity) from pseudo-science (e.g., Freud's psychoanalysis). Or how much of the current biomedical literature are not reports of "puzzle-solving" - as described by Kuhn? (See for example, the set of three papers reviewed in section 16 of this report.) It is also worth noting that the views of both men are not altogether original: Popper's views draw heavily upon those of the nineteenth century polymath Whewell, and Kuhn's were strongly shaped by his early reading the pathologist Fleck. Both were to some extent anticipated by the physicist C.S. Peirce. And the distinguished chemist John Platt arrived, independently, at a point of view quite similar to Popper's. Platt (1964), combining the earlier precepts of Francis Bacon with the more recent examples of scientific achievement in molecular biology and nuclear physics, recommends that every scientific study be evaluated on the basis of whether or not its findings refute any accepted hypothesis on the matter at issue (his so-called "method of strong inference").

We may note at once that the concepts, methods, and criteria by which the answers to the last three questions were obtained and technically validated are, for the most part, those provided by currently accepted practices in statistical modelling, namely, the generalized linear models, regression diagnostics, and Bayesian regression, as described in, for example, the work of Belsley, Kuh and Welch (1980), Chatterjee and Hadi (1988), Cook and Weisberg (1982), Dobson (1983), Gilchrist (1984), Leamer (1978), McCullagh and Nelder (1983, 1989), Montgomery and Peck (1982), Theil (1971), and Zellner (1971).

The philosophical - and sociological - contexts in which, as it seems to us, these answers may be best understood, interpreted, scientifically validated, and further exploited, is that provided by the works of Bacon (1620/1960), Fleck (1935/1979), Jeffreys (1957, 1961), Kahneman, Slovic, and Tversky (1982), Kuhn (1970a, 1977), Nisbett and Ross (1980), Peirce (1955), Popper (1965a, 1965b), Ross and Lepper (1980), Whewell (1860/1971), and Ziman (1968).

In his Presidential Address, "The Importance of Statisticians," at the 1986 Annual Meetings of the American Statistical Association, Donald Marquardt, echoing the remarks of Sir Maurice Kendall of two decades earlier that were quoted above and in Annex I, noted that, "Statistics is the discipline responsible for studying the scientific method with the greatest intensity and for providing in-depth expertise to other disciplines." and, moreover, "The fundamental role of statisticians is to be purveyors of the scientific method." Dr. Marquardt noted further that among the professional competitors to statisticians were included the physicists. Now it is the case, as the poet Robert Burns has remarked, that it is often useful, "to see oursel's as others see us." The findings of this task group report, together with the views of Marquardt - and of Burns - suggest that the medical physicists must consider their larger professional role to be, "purveyors of the scientific method" to the medical profession. They must become scientific generalists who can, "practice science - not a particular science" (Bode, Mostellar, Tukey and Winsor, 1949).

2. "What do we believe?"

"There is nothing so absurd or impractical that it has not been asserted by one philosopher or another."

Rene Descartes, 1643[2]

"We now realize that knowledge in science is in large part conventional. New facts are uncovered and accepted as 'true' within a system of shared conventions by which members of a scientific discipline seek to interpret their observations of nature."

S. Jasanoff, 1989

"These [paradigms] I take to be universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners."

T. Kuhn, 1970

2.1 Received beliefs

Much of what we believe - concepts, estimates, inferences, methods, and criteria - in radiation dose-response is conveyed by the following assertions which are extracted from the studies listed in Table 1. Upon (secondary) analysis of the data on which they were reported to have been based, several of these statements were found to represent, to borrow two quite apt phrases from earlier writers, the "... illicit generalizations and hasty conclusions" (F. Bacon, 1620), and "... the imposing delusions of received theories" (W. Whewell, 1856).

2.1.1. Radiation carcinogenesis and neoplasia

a) "The data strongly suggest that leukemia risk is increased by exposure to gamma radiation. ... they suggest that risk from low-LET radiation may be estimated by using the gamma-dose coefficients in the fitted linear-quadratic gamma linear neutron (LQ-L) model." (BEIR III, 1980).

b) "Breast-cancer data offer little support for a dose-response model with strong upward curvature in D_γ. The dose-response curves for mammary tumors in female rats given total-body x and gamma irradiation tend to be linear. Functions of D_γ and D_N fitted to the breast-cancer incidence rates for Hiroshima and Nagasaki, standardized to the age distribution of the combined cities, suggested a relationship linear in both D_γ and D_N ... Linear-model coefficients for D_γ and D_N did not differ significantly, and the RBE values most consistent with the data gave linear-model risk estimates that differed only slightly from those obtained with the assumption of an RBE of 1. Accordingly, the following analyses of the LSS sample data do not distinguish between the gamma and neutron components of breast-tissue dose." (BEIR III, 1980).

c) "332. For sparsely-ionizing radiation, most curves are upward concave and may be fitted by linear-quadratic or quadratic (leukemia in CBA/H mice) models, but in some cases approximate linearity may apply. This is so with mammary fibroadenoma and carcinoma in Sprague-Dawley ... rats." (UNSCEAR, 1986).

d) "Essentially without exception in eukaryotic systems, the RBE of high-LET radiation is a strong function of dose, increasing as the dose decreases."(NCRP 64, 1980).

e) "In these animals, total body x-ray or ^{60}Co gamma-ray irradiation at 1-2 months of age advances the onset of tumors; and the incidence of total tumors scored within one year after exposure increases as a linear function of the exposure from 25-400R with x-rays (Bond et al, 1960) and 16-250R with ^{60}Co gamma rays (Shellabarger et al, 1969)." (NCRP 64, 1980).

f) "When the percent of rats with mammary neoplasia is plotted against dose ... for the present experiment ... and a straight line calculated by means of a least squares fit, as done in Fig. 5, two things are apparent ... there is a fairly good direct relationship of dose and response. As the dose is doubled the response doubles approximately. Note that the regression lines are calculated excluding the zero dose and only up to 200 or 250R. Even so, the line calculated extrapolates closely to the experimentally determined response at zero dose." (Shellabarger et al, 1969).

g) "457. ... from such comparisons, the pattern seems to emerge that the shape of dose response relationships may be basically similar in various species. 458. Thus, the linear non-threshold type of dose-response relationship for female mammary carcinoma induced by x-rays finds its counterpart in the results of studies on at least four strains of rats: (a) the Sprague-Dawley females irradiated externally by x-rays and/or gamma rays ..." (UNSCEAR, 1986).

h) "The relative biological effectiveness of the 35 MeV neutrons compared to γ-rays for breast tumor induction was found to be 4.3 based on the slopes of the linear responses to both types of radiation. With the best-fit curves for both neutrons and γ-rays, the relative biological effectiveness increased from 5.0 at 40 rads of neutrons to 13.8 at 2.5 rads of neutrons." (Montour et al, 1977).

i) "Although not all these objections seemed cogent to all members of the Committee, it was agreed that further modification of the LQ-L model [of cancer sans leukemia mortality] would be desirable and that the leukemia experience might provide a reasonable, if arbitrary, guide. This model, denoted LQ-L, is

$$F(D_\gamma, D_n) = \alpha_0 + \alpha_1(D_\gamma + 0.0086D_\gamma^{\,2} + 27.8\ D_n) \qquad (V-10)$$

That is, not only is the ratio α_2/α_1 in Equation V-3 fixed at the leukemia value of 0.0086, but the neutron RBE is implicit in the model. This further change yields a more stable estimate of the coefficient, $\alpha_1 = 1.40 \pm 0.38$." (BEIR III, 1980).

j) "Accordingly, the Working Group has adopted linearity for breast and thyroid cancer and the BEIR III linear-quadratic model for all other sites, for PC calculations involving exposure to low-LET radiation. The BEIR III linear-quadratic model estimates excess cancer risk, following a radiation exposure of short duration as

$$\text{Excess} = e_{LQ}x(D + D^2/116)$$

where e_{LQ} is a site-specific coefficient depending upon age at exposure and sex, and D is the radiation dose in rad ... The linear model, of course, expresses excess risk as

$$\text{Excess} = e_L \text{ x } D$$

The so-called cross-over dose of 116 rad in the formulation of the BEIR III linear-quadratic model, which specifies the degree of curvature in the graph of risk as a function of dose, was originally determined from the Japanese leukemia data ... but is also consistent with a number of other radiobiologic endpoints. ... Since no other human cancer data are adequate for calculating cross-over doses for individual sites other than the thyroid gland and female breast, the value of 116 rad has been assumed to apply to all sites for which a linear-quadratic dose-response model for low-LET radiation is considered appropriate for low-LET radiation (i.e., all cancers other than those of the thyroid and female breast)." (NIH Ad Hoc Working Group, 1985).

2.1.2. Radiation Mutagenesis

a) "The numerical values and the variability in the estimates of α, β, and α/β in such favorable material as Tradescantia are illustrated in Table 5.1."(NCRP 64, 1980).

b) "The apparent curvilinearity of the x-ray line can be approximated by two straight-line segments, one with slope + 1.4 from about 5 to 100 rads, the other with slope + 1 from 0.25 to 6.0 rads. A t-test indicates that those slopes are significantly different ($P < 0.01$). However, a more meaningful interpretation is that the entire ascending portion of the x-ray curve can be fit as a sum of a linear and a quadratic dose term." (Sparrow et al, 1972).

c) "324. The induction of somatic mutations in the stamen hair of the plant Tradescantia was studied over a wide range of doses, dose rates, and radiation qualities. This system provides direct and extensive information at very low doses of x and gamma rays and of neutrons. The relevant dose-response relationships fit very well a linear-quadratic model for low-LET radiation and a linear model for neutrons." (UNSCEAR, 1986).

d) "161. The data obtained on plant cells of the species Tradescantia are very valuable because: (a) mutations can be scored down to 3 mGy of x-ray (and correspondingly lower doses of neutrons); (b) the amount of data on the effects of dose, dose-rate, and radiation quality is far better than for any other cellular system." (UNSCEAR, 1986).

2.1.3. Radiation Lethality (cell survival)

a) "Using the linear quadratic (L-Q) model, which provides an overall adequate fitting especially in the low-dose range ..." (Fertil and Malaise, 1982).

b) "It is relatively easy (using a programmable pocket calculator) to determine the parameters

α and β by performing a least squares regression of the function $y = -\alpha D - \beta D^2$ from the natural logarithm of the survival fractions. This method is comparable to a weighting procedure leading to a similar contribution of all experimental points." (Fertil and Malaise, 1981).

c) "For each cell line and with each model, we obtained a survival curve characterized by the fitted parameters. The experimental fluctuations and the quality of the fit were expressed by both the variances and the covariance(s), linked to these parameters ... the quality of fitting represented by a 95% confidence ellipse or ellipsoid ..." (Fertil et al, 1980).

d) "The use of the LQ model was solely dictated by its outstanding descriptive properties." (Fertil and Malaise, 1985).

2.1.4. Radiation toxicity

a) "The dose-response formulae that could be used, with statistical legitimacy, include (1) the linear-quadratic model; ... For present purposes the LQ model offers (a) valid approximations for all doses likely to be used in radiotherapy, (b) only two parameters to be determined, and (c) considerable convenience in practical applications."; "If a series of fractionated schedules have been used,so that isoeffect total doses are known for a set of dose per fraction, the ratio (α/β) can be found simply by plotting reciprocal total dose against dose per fraction." ... "... plotting [D^{-1} vs D/N] yields a line whose straightness is a test of the validity of the [$\alpha D + \beta D^2/N$] formula." ... " d + d2 provides a good fit to experimental data in the shoulder region, at least up to several grays (per fraction) ..." "[the] $\alpha d + \beta D^2$ model [is believed to be] ... valid for doses per fraction in the radiotherapy range, i.e., up to about 10 Gy ..." (Here d = D/N, the dose per fraction).

"I shall use the $\alpha d + \beta d^2$ model because I believe it to be ... valid for doses per fraction in the radiotherapy range, i.e., up to about 10 Gy ..." (Fowler, 1984).

b) "In any case, LQ is simply a low-dose approximation ..." (Fowler, 1989).

c) "If the LQ model applies, the reciprocal total isoeffect dose is related linearly to the size of the dose per fraction, and α/β is the ratio of the vertical intercept to the slope of the resulting line." (Tucker, 1984).

d) "Mathematically, the effect of this time factor can be included in the isoeffect formula [$D(\alpha + \beta D/N) = E$ = constant effect level] by subtracting a function f(T) ... The data suggest that a reasonable form for f is the linear function $f(T) = \gamma T$ for some constant γ over the dose per fraction and time range specified." (Travis and Tucker, 1987).

e) "For the abscissa, the slope s_n of each quantal curve in Fig. 5 was estimated by eye." (Tucker and Thames, 1983).

f) "Assuming that the average number of doses used to obtain each quantal curve was $n_D = 3$ (as indicated in Fig. 5), and that $n_A = 10$ animals were irradiated per dose, the value of the parameter k for those data is approximately k = 0.175 (c.f. Table 1)." (Tucker and Thames, 1983).

g) "There is now a considerable body of evidence to suggest that the LQ model may be of wider validity in fractionated radiotherapy than the well-known Ellis formula and its derivatives (i.e., the TDF and CRE equations) and some centres are already using the newer formalism in place of more traditional methods (Fowler, 1984)." "... Following the reasoning of Barendsen (1982) and Thames et al (1982), we may further infer that biological effect (E) in irradiated tissues is uniquely determined by the surviving fraction of target cells. The level of effect is related to cell survival by: E = -log (surviving fraction), i.e., using eqn. (1) we have,

$$E = \alpha D + \beta D^2 \qquad (2)$$

If eqn. (2) is rewritten as

$$E/\alpha = D[1 + D(\beta/\alpha)]$$

then, since α is a constant, the effect of the radiation is seen to be dependent on the product of the dose D and a function of D, the latter depending upon particular values of (α/β)." (Dale, 1985).

h) "Although the reciprocal total dose of Fe-plot (Douglas & Fowler, 1976) is not the most accurate way to calculate α/β, it is the easiest method and gives a fairly accurate value if the data are good, but only an optimistic estimate of the error range." (Fowler, 1989).

i) "The object of clinical radiation therapy is to obtain, for each patient, the maximum probability of cancer cure while minimizing the likelihood of significant normal tissue damage." (R. Yaes, 1988).

j) "For late-reacting tissues the time factor is zero, and it should be. There is no disadvantage in the LQ formula having no time factor for late reacting issues (sic); indeed it is an advantage. Since its main application is in the calculation of isoeffect doses to avoid over-dosing late-reacting tissues, the LQ formula as used above [$E/\alpha = nd(1+d(\beta/\alpha))$ = constant] has been and should remain sufficient, ..." ... [Where it is required, a] "time factor can be added simply by multiplying the surviving fraction S by an exponential factor $e^{\gamma T}$ which of course works in the reverse direction to the killing effect of radiation:

$$S = \exp[-n(\alpha d + \beta d^2) + \gamma T] \qquad (7)$$

so that the total effect E is given by:

$$E = n(\alpha d + \beta d^2) - \gamma T$$

and the biological effective dose is given by:

$$\frac{E}{\alpha} = nd\left(1 + \frac{1}{\alpha/\beta}\right)\frac{\gamma T}{\alpha} \text{ (sic)}$$

...

"It may be said that the LQ model loses its innocence when a time factor is added." (Fowler, 1989).

k) "Let -E be the natural logarithm of the fraction of surviving target cells corresponding to this degree of injury. ... It follows easily from the LQ model that there is a linear relationship between the dose per fraction $x_n = D_n/n$ and the reciprocal of the total isoeffect dose:

$$(1/D_n) = (\alpha/E + \beta/E)\, x_n \qquad (1)$$

Estimates of the intercept α/E and the slope β/E can be obtained by the linear regression of $1/D_n$ against x_n (Fig. 2). Since, in general, the value of E is not known, the parameters α and β cannot be determined, but the ratio $\alpha/\beta = (\alpha/E)/(\beta/E)$ can be calculated independent of the chosen level of effect." (Tucker and Thames, 1983).

l) "The basic assumptions in applying the LQ model for predicting isoeffect doses in different fractionation schedules may be summarized (D28):
a) the biological effect E of a dose d is determined uniquely by the surviving fraction S of a target cell population. E is adequately approximated by the relation:

$$E = -\log S = \alpha d + \beta d^2 \qquad (1)$$

b) equal reduction in log survival is obtained after each fraction, c) repair is complete, and d) proliferation is negligible. It follows directly from these assumptions that the effect of n equally sized fractions is given by the multifraction LQ model

$$E = n\alpha d + n\beta d^2 = \alpha D + \beta dD$$

where D is the total dose given." (Bentzen et al, 1987).

m) "The objective is to maximize the probability of tumor control without incurring an unacceptably high probability of organ damage."

...

"The α/β ratios for late reacting tissues could be estimated from the available clinical isoeffect data and from animal experimental data by using the F_e-plot of Douglas and Fowler (4)." (R. Yaes, 1989)

n) "Even though it is true that once normal tissue tolerance is reached the treatment has to be stopped whether the tumor lethal dose is reached or not it is absolutely necessary to verify that the tumor damage is extensive enough to have beneficial effect for the patient. ... the concepts of NSD/CRE, TDF and BIR are centered around the normal tissue tolerance and not on tumor lethal effect, in the case of commonly used treatment schedules, adequate tumor lethal effects are achieved simultaneously. In extreme fractionation it may not be so."

"In the case of malignant cells, these may be represented by $N^{-p}T^{-q}$ where p and q are less than 0.24 and 0.11, respectively. The tumor significant dose should therefore be represented by TSD

$= DN^{-p}T^{-q}$. In the trial expressions for TSD, we used the following combinations of p and q: (i) p = 0.24 and q = 0; (ii) p = 0.22 and q = 0.02; (iii) p = 0.20 and q = 0.04; (iv) p = 0.16 and q = 0.08; (v) p = 0.13 and q = 0.11; (vi) p = 0 and q = 0.22."

...

"The values of p and q were evaluated from patient data from the following two sources: (i) from the cases of patients treated for cancer of the cervix uteri from the period 1978 to 1980 at the V.N. Cancer Centre, G. Juppuswaney Naidu Memorial Hospital, Coimbatore, and (ii) the cases reported by the British Institute of Radiology Study Group under a scheme of comparison of 3 fraction/week treatments with 5 fraction/week treatment for cancer in laryrgo-pharynx region."

...

"From the above, it was concluded that the value of p lies between 0.16 and 0.20 and corresponding q lies between 0.08 to 0.04."

...

"In order to evaluate the malignant tissue damage a concept of tumor significant dose has been defined in this work. The working formula for the TSD is

$TSD = DN^{-0.18}T^{-0.06}$

...

Supe et al, 1983

o) "In his analysis of factors determining success or failure in radiotherapy of cancers of the skin and lip von Essen derived best-fitting exponents in the empirical power function for tumor volume and skin area. The exponents were shown to differ in both magnitude and sign in the two situations. Since normal tissue tolerance is inversely related to field size the exponent for normal tissue has a negative sign. Its magnitude was estimated to be $\upsilon = -0.27$. Since the effective tumor dose varies directly with tumor size, the exponent for tumors must be positive and the best estimate for this parameter was given by $\upsilon = +0.14$. von Essen incorporated those parameters into isoeffect formulae for tumor control and normal tissue tolerance, and developed a complex iso-response grid as a guide to management." (Cohen, 1982).

p) "Superimposition of these two sets of lines with different slopes illustrates the proposed model (Fig. 10). ... In addition, only two of the isoeffect lines have been derived directly from the data. The construction of all the other curves parallel to these two originally derived curves is then, hypothetical." (von Essen, 1960).

We must here note, in passing, a most important issue to which we shall repeatedly recur: Despite the relentless production of large numbers of papers on LQ models of radiation effects over the past ten or so years, statistically adequate evidence that the LQ model is valid for any radiation response (other than chromosome aberrations and, just barely, the LSS (DS86) data on leukemia mortality as reported by the BEIR V committee in 1990), even - or especially - as a low-dose approximation, and, for the multifraction LQ model of radiation toxicity, with or without a "time factor", has rarely been published. Much of the present report has to do with the evaluation of the concepts, methods, and criteria from which unfounded beliefs are conceived and developed and by which they are subsequently maintained.

2.1.5. A note on some of the criteria for the selection of studies for review by the Task Group

Our first question concerning dose-response phenomena - "What do we believe?" - was, of course, answered by simply compiling an eclectic selection of papers describing the several models, estimates, etc. that have been chosen to convey the received wisdom on the respective roles of dose, time, fractionation, volume, etc., in eliciting, and in modulating, the several responses of irradiated cells, tissues, organs and individuals, together with the empirical and a priori evidence adduced for each therein. We also selected several papers that described the received methods and criteria by which those models were constructed, selected, and deployed.

In making this selection of papers and reviews for evaluation that are given in Table 1, considerable weight was given to the citation frequency of a paper, as found in the Science Citation Index (SCI) since several studies have disclosed a strong correlation between the "importance" of a paper (as determined by a panel of judges) and the citation frequency (Virgo, 1977). However, it must be noted at once that the citation frequency of a paper has been shown to be only weakly - albeit positively - correlated with the quality - methodological rigor - of the paper. In fact, the latter correlation is too weak for the citation frequency to be a reliable indicator of an individual paper's methodological rigor (Bruer, 1982).

This proved to be true of the studies examined in this report also: The evidence for the conclusions presented in the more frequently cited papers of Table 1 was usually no more convincing than that presented in the least frequently cited paper (Annex IV, part 3). Indeed, the substantive conclusions of several of the more frequently cited papers were found to be exceedingly difficult to defend and many could be rejected on the basis of the data that were offered in their support (Annex II, part 3; Annex III, parts 3-6; Annex IV, part 4).

Nonetheless, this method of selection should provide a reliable account of "What we believe"; that is, of what the peer group has identified as legitimate problems and has defined as acceptable solutions (methods and criteria) thereto, i.e., it provides an account of the paradigm that informs and guides research praxis in the radiobiology peer-group. ("The assumption is that highly cited papers have a major impact on research; numerous studies suggest that this is a valid assumption." F. Narin and J. Frame, 1989.)

As well as the citation frequency of the paper in which it was reported, additional weight was attached to a study if it were discussed - or even mentioned - in an authoritative review of the literature, such as those of Fowler (1984, 1989) on the LQ model, Cohen (1982), Withers, et al (1988) on volume effects, and the institutional reviews of NAS/NRC, NCRP, and UNSCEAR on the biological effects of low doses of ionizing radiation. Ziman (1968) has briefly discussed the role of the review article: "The author is expected to read all the papers on the subject, give a brief account of their findings and relate them to one another, noting agreement and contradictions ..." "The function ... is to give an explicit account of the current consensus in some particular field. ... From such a review, one can learn the general credibility status of each paper and the weight of its contribution to the advancement of the subject." (Note that Zieman's description of the current role of the review article reveals it to be a largely qualitative and informal (and inadequate) meta-analysis. See section 1.1 above.)

We note that these criteria for the selection of studies in Table 1 are consistent with the so-called Frye "general acceptance test" for the admissibility of a scientific principle as probative evidence in a court of law (Frye vs United States, 1923. See Black, 1988a, b.) Both require evidence of acceptance by the peer-group. See part 13 below.

One other criterion for the selection of the studies listed in Table 1 is that the primary data and/or a complete description of the computational methods deployed by which the respective investigators obtained the published results for the study must, of course, be available to Task Group 1. This criterion also confined the evaluation of those authoritative reviews listed in Table 1 to those sections which referred to such studies.

2.2 Publication bias

Critical review of a selection of published papers provides sound information on the nature of the received paradigms that inform both the thinking and practices of the peer-group in a given field. However, basing any review on only the published literature of the field exposes the rival hypotheses, estimates, and inferences that may be constructed by rival methods and criteria from the same data from which the received hypotheses, estimates, and inferences were constructed, to the risk of being weakened by the effects of a version of the well-known publication bias that results when not all investigations of a given issue are equally likely to appear in published journals. This is because the designs by which much of the published data was acquired in dose-response experiments are encumbered by the severe weaknesses that we will examine in parts 5-11 of this report. For, as Blakemore et al (1963) have remarked, "Carefully designed experiments are necessary ... there are no fitting techniques which can overcome the deficiencies of poorly designed experiments." (echoing Fisher's comments of a generation earlier: "If the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too.") For example, it is quite difficult to determine, "What must be the case", say, for the separate roles of fractionation and protraction in modulating the biological effects of radiation dose, owing to the high degree of correlation, in the joint distribution of the observations, of the number of fractions, N, and the number of days, T, in the data of all published experiments that we reviewed - as well as, of course, in all non-experimental data bearing on the issue. On the other hand, it is, of course, much easier to determine what is not the case in this matter - even from flawed experiments. And, in the case of the experiments in radiation carcinogenesis that we examined, two simple transformations of the data permit one to say, with acceptable precision, what the case must be with respect to the issues of thresholds in the dose-response curve and the dose-dependence of the neutron RBE.

However, the issue of publication bias is a still more general and insistent problem and some further definitions and other appropriate illustrations of the issue must be introduced at this point: "As a working definition, publication bias is the bias produced when the decision to publish is influenced by the results of the study" (Berlin et al, 1989). More precisely, publication bias is introduced when both the decision (by the investigator(s)) to submit manuscripts for publication and the decision to accept or reject (by the Editor) are influenced by the statistical significance or "positivity" of the results. For example, in the case of reviews of randomized controlled clinical trials (RCTs) there is considerable evidence that, "the published clinical trials may be biased in favor of 'significant' or 'promising' results. Clinical trials which fail to show any treatment difference are less likely to be published owing to investigators not writing up the results or owing to journals declining to publish such 'uninteresting' information" (Simes, 1987). One notes that there is also the possibility of "publication bias via suppressed criticism". This occurs when a paper that demonstrates that a previous study is in error due to "... wrong reporting of a statistical analysis ..." is rejected for publications: "Rejection of a paper by a journal seems to be an infrequent cause of publication bias but is not unwillingness to acknowledge mistakes another form of bias, and one on which the editors of biomedical journals should take a definite position if the confidence of their readers is to be maintained?" (G. Landi and A. Ciccone, 1993. Emphasis added). Two additional remarks will further illuminate the issue: a) "In many fields of research and in some less respected journals, two kinds of studies have a smaller chance of being published: those which do not reject a null hypothesis and/or studies whose findings are not coherent with current prevailing paradigms or a body of knowledge ... Consequently, a portion of the complete body of information, in the form of unpublished studies, remains 'in file drawers'." J. Jenicek, 1989. b) "Two types of bias in the published literature must concern a reviewer. First, because authors and journal editors tend to report significant findings, a review limited to published studies will tend to over-estimate the effect size." ... "Second, another form of publication bias, the confirmation bias, tends to emphasize and believe experiences that support one's views and to ignore or discredit those that do not ... Consequently, new or unpopular data tend also to be under-reported in the published literature." S. Thacker, 1988.

Three recent and egregious instances of "confirmation bias" can be found in the experiences of three of the foremost workers in the nascourt field of non-linear dynamics: David Ruelle, Boris Belousov, and Yoshisuke Ueda. David Ruelle's revolutionary 1971 work on deterministic chaos in dissipative systems (which eventually prevailed over the then-received Landau-Hopf model of the transition from laminar to turbulent fluid flow) was repeatedly rejected for publication when it was first reported. Similarly, Boris Belousov's equally revolutionary 1951 work on the sister phenomenon of self-organization (a simulation of the Kreb's cycle in which chemical oscillations were first remarked) was initially rejected. (It eventually appeared as an obscure paper in the proceedings of a symposium on radiation medicine in 1959). See Ruelle 1991 and Coveney and Highfield, 1991. Ueda first observed chaotic behaviour in a simple Duffing's oscillator in 1961. His attempts to get his observations published in 1962 and again in 1971 were unsuccessful. "Not until 1978 did his famous 'Japanese Attractor' paper get published ... . The moral of this story is clear: Even in one of the most advanced laboratories in non-linear circuits, chaotic dynamics were rejected because they did not fit in with the mathematical theories of the times." See Moon, 1992.

Recent studies of the issue of publication bias disclose that it is widespread and in some studies may be quite large. When that is the case it may severely degrade the estimates and inferences obtained by a meta-analysis. One of the more striking examinations of the issue described an assessment of publication bias using a sample of controlled clinical trials. Three primary endpoints were examined: overall patient survival, disease-free survival, and tumor response rate. There were "... striking trends for each endpoint, with small studies appearing to possess large treatment effects and large studies possessing relatively small treatment effects. It is believed that these differences are primarily due to publication bias. The bias is very large: Absolute differences observed were 41% for overall survival, 79% for disease-free survival, and 17% for response rates ... An implication of this study is that the results of small published studies are typically unreliable, even taking into account the fact that such trials are imprecise due to sampling variation" (Berlin, et al, 1989). In a "worst case" view of the extent of the problem - the so-called "file-drawer problem" - the journals are filled with the 5% of the studies that show Type I errors while the file-drawers back at the laboratories are filled with the 95% of the studies that show "non-significant" results (Rosenthal, 1979).

The issue of publication bias first arose in an examination of the now well-documented, "prejudice against the null hypothesis" (Rosenthal, 1979) that sharply reduces the rates of both submission and publication of any negative results, e.g., results which are not statistically significant or, as remarked above, which controvert the conventional wisdom on a given matter. Kuhn has observed that, "Normal Science does not seek anomalies and when it is successful finds none." (1970a). Since it has been estimated that a study may be more than ten times as likely to be submitted and accepted for publication if the findings are positive rather than negative. Kuhn may be correct. These aspects of publication bias raise a more fundamental question - and one with serious implications for clinical radiation biology: The fact that the present task group report seems to be the first publication to disclose the presence of weaknesses in current radiobiological praxis, for which many are as obvious as they are egregious, suggests that publication bias may have suppressed earlier similar findings - in an effort to maintain the consensus that defines the practice of Kuhn's "Normal Science."

2.3 Do scientists read scientific papers?

"... scientists have a strong urge to write papers but only a relatively mild one to read them."

D. J. Price, 1963

"Today, publication of papers on clinical investigations promptly affects the current practice of medicine."

M. Zelen, 1983

The "... weighted average review time (adjusted for number of reviews) was 2.4 hours."

A. Yankauer, 1990

"For every person who reads the whole of the text of a scientific paper, twenty read through the Abstract of the paper, and about five hundred read the title of the paper and stop there. Most papers are read by title alone."

G. A. Kerkut, 1983

One striking feature of scientific practice that was disclosed by the review of the literature on which the present report is based is that the literature which is cited is not always read very carefully - cited but not read. Or, perhaps better, read without result. Since some of the literature that is thus cited is flawed, this effect - or practice - contributes to another kind of publication bias. For although - as Ehrenberg has remarked - saying the same thing twice does not make it more true, it nevertheless does make it more "believable", as an effect of a kind of induction by enumeration, an argument from a "muster of instances" (Bacon, 1620). In particular, the appearance of a citation of a previous study is evidence that some degree of that consensus, which is a potent warrant for the scientific "truth" of any proposition, has been achieved, on the matter at issue.

The evidence for this is that many of the study weaknesses that we have identified, classified, and provided measures of their respective effects on the estimates and inferences that comprise the study findings, were apparent to even a casual reading of the reports examined. For example, 1) it is clear from the Shellabarger et al, 1969 report on radiation-induced mammary neoplasia that only by deletion of the observations at the lowest and highest doses will the data "fit" their linear model. 2) It is obvious from Table V-8 of the BEIR III (1980) report that the LQ-L model over-fits the leukemia incidence data. 3) It is evident from several F_e-plots that both the goodness-of-fit and the estimate of the α/β ratio are dominated by a single observation at N=1 that is well-beyond the stipulated range of validity of the LQ model ($0 < D/N < 10$ Gy). 4) It is obvious from von Essen's 1960 paper on "volume effects" that his 1963 estimates of the exponents 0.075 (tumor) and -0.16 (skin) for the area, or volume, of irradiated tissues are based on hypothetical isoeffect doses. 5) It is clear that the exponents 0.14 and 0.27 in von Essen's 1960 paper on "volume effects" do refer to the elapsed time of the treatment schedule and not to the area or volume of irradiated tissues (as reported in one authoritative review of that paper). 6) It is clear from Fig. 8 of the Tucker and Thames (1983) paper that the point and interval estimates of the flexure dose are dominated by a) the single observation at N=1 (beyond the stipulated range of validity of the LQ model and b) the implausible constraint that forces the regression line through the origin - the constraint requires that the slope, sn, of the (conditional) dose-response curve at P = 0.50 be equal to zero and that the ED50 dose to be negative (See Annex IV, part 4). 7) It is evident from Fig. 5 of the Tucker and Thames (1983) paper that these dose-response data are badly flawed since a) most of the data (63%) resides in the extreme responses $r_i = 0$ or $r_i = n_i$, which although $n_i > 1$ provide no more information than if $n_i = 1$. ("The extreme situation is that, in every batch tested, either all members respond or all fail to respond, so that the evidence from a batch is no more reliable than that from an individual." Finney, 1971b), b) the numbers of levels of dose at each of the seven levels of fractions and time varies between 2 and 4, c) much of the data (47%) lie well beyond the stipulated range of validity of the LQ model, d) the goodness-of-fit and parameter estimates of any model will be dominated by the single dose observations - which lie well beyond the stipulated range of validity of the LQ model. 8) It is apparent at once that the ratio α/β of the LQ model can be unique only to within a multiplicative constant and hence is unsuitable as a discriminator of radiation effects. 9) It is obvious that the version of the multifraction LQ model that is commonly deployed (e.g., Tucker and Thames 1983; Bentzen et al, 1987): $D^{-1} = (\alpha/E) + (\beta/E)D/N$ where $E = -\ln S$, has little empirical content since neither the identity of the cell-population to which S, the survival, refers nor the level of S itself can be specified - or even defined. And so forth.

It would seem that the earlier observations of Price (1963) as well as the more recent findings of Kerkut (1983) hold good for much of the current practice in the field of dose-response modelling. Our review also suggests that the findings of two recent (circa 1990) studies, which disclose that the (peer) reviewers for medical journals spend only about two hours per manuscript, are not too wide of the mark in this field.

It has been recently (1990) been reported in Science that more than half of all papers published in major journals are never cited after publication. This finding has raised several issues of the validity and relevance of the bulk of the research that is currently funded. Our findings not only reinforce an earlier proposition that the frequency of citation is not a reliable measure of the validity and relevance of a study but it is not even a reliable index of whether the study had even been read; that is, read with result.

David Horrobin has remarked that, "The ultimate aim of peer review in biomedical science cannot be different from the ultimate aim of medicine ... 'to cure sometimes, to relieve often, to comfort always'." (Horrobin, 1990). To which one must add, "First do no harm." (See section 10. "Statistical methods and the ethics of scientific enquiry".)

3. Models of radiation dose-response

"Understanding what is going on around us is equivalent to building models and confronting them with observations."

G. Nicolis and I. Prigogine, 1989

"All models are wrong but some are useful."

G. E. P. Box, 1979

"One cannot escape the feeling that these mathematical formulas have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverer, that we get more out of them than was originally put into them."

Heinrich Hertz, 1883

"... a good model should not only provide a structural description at the phenomenological level, but also admit a reductionist explanation at the microlevel."

E. C. Zeeman, 1982

It is obviously central to the purposes of the present report to have in hand concise definitions of dose-response models and their several functions. Thus, "A model is a formalized expression of a theory or causal situation which is regarded as having generated the observed data. In statistical analysis the model is generally expressed in symbols ..." Although they vary from the heuristic to the ontological, "... all models have similar functions. Among other things they supply the group with preferred or permissible analogies and metaphors. By doing so they help to determine what will be accepted as an explanation and as a puzzle-solution; conversely, they assist in the determination of the roster of unsolved puzzles and in the evaluation of the importance of each." (T. Kuhn, 1977)

In particular, the regression models of dose-response (and isoeffect) achieve what the physicist Mach (and the statistician Pearson) have described as economy of thought: the n observations which comprise the data are replaced by the k << n parameters of the model which summarizes them and by means of which they can be reproduced (with greater or lesser fidelity). One of the basic criteria of the validity of the model is the degree of fidelity with which the model can reproduce the observations.

3.1 Types of scientific and statistical models

'Scientific models are of two general kinds. 1) Phenomenological theories (or "black box" models) are concerned with directly observable variables. Black box models focus on the behaviour of a system rather than on its structure; they include relations between the inputs to a system and the outputs from the system. 2) The second kind of model, representing theories ... goes somewhat deeper. These models include 'speculation on what goes on in the innermost recesses of reality'" (M. Bunge, 1979). These have been defined as deep models by Bross (1970): "The goal of a 'deep model' is to predict the observable events from the parameters that characterize the events at the microscopic level." (Lionel Cohen's cell population kinetic (CPK) model is an example of a "deep model". S. Gould (and others) have stressed the role of metaphor in the genesis and development of scientific hypotheses. Cell survival curves have provided the metaphor for Fowler's LQ models as well as Cohen's CPK model. Bioassay and the Schwarzchield effects have provided the metaphors for the so-called Power law models of radiation effects.)

There are three main types of statistical models: 1) the ontological, or functional, model, sometimes also referred to as the mechanistic model; 2) the control model; 3) the heuristic, phenomenological, or empirical, model. If the true mechanism of the process that generates the observed response is known, a priori, then the functional relationship that subsists between the response and predictor variable is known and can often be exploited by the investigator to understand, interpret, control, and predict the response over a rather wide range of the predictor variables with considerable accuracy and precision.

However, even if the functional relationship is known completely there are often circumstances in which the functional model is not useful for achieving control of the response variable, since there are often one or two variables that may profoundly affect the level of response but are not under control of the investigator although these variables can be identified and the functional relationship of these variables to the response may be well-known. However, it is sometimes possible to construct a useful model of the response in which all of the predictor variables are under control of the investigator. Such models often include surrogates, or proxies, for the "explanatory" variables that are not accessible to the investigators, or cannot be identified. Indeed, as Fisher (1966) has observed, "... in the state of knowledge or ignorance in which genuine research intended to advance knowledge, has to be carried on ... we are usually ignorant which, out of innumerable possible factors, may prove ultimately to be the most important, though we may have strong presuppositions that some few of them are particularly worthy of study. We have usually no knowledge that any one factor will exert its effects independently of all others that can be varied, or that its effects are particularly simply related to variations in these other factors. On the contrary, when factors are chosen for investigation, it is not because we anticipate that the laws of nature can be expressed with any particular simplicity in terms of these variables, but because they are variables which can be controlled or measured with comparative ease."

When neither mechanistic nor control models that are useful to the purposes at hand can be constructed, it may be possible for the investigator to construct an empirical, or phenomenological, model, which, though it may in some respects be uninterpretable, or even unrealistic, will reproduce the main features of the process sufficiently well to enable the investigator to exploit it to make useful predictions. In certain circumstances these empirical, or predictive, models serve heuristic purposes as well, leading to valid insights into important aspects of the process that generates the observed set of responses.

However, the control and empirical models are, strictly speaking, useful only for interpolation over the range of the data. The mechanistic model provides a better basis for accurate and precise extrapolations beyond that range because it is based on at least partially validated understanding of the process that generates the observations; it represents more than a curve or surface which merely (albeit perhaps adequately) graduates the observations. Therefore, the bias in the estimated response is less. Moreover, since the mechanistic model is (usually) more parsimonious than the empirical model, less of the random error in the observations will be transmitted to the estimated response and hence the variance is also less. The average variance of the estimated response over a sample of n observations is proportional to k/n where k is the number of free parameters in the model. (Montgomery and Peck, 1982)

But, "... as we move in the space of the experimental variables, the mechanism may change or estimation errors may become serious, so unchecked extrapolation is never safe," (Box, Hunter, and Hunter, 1978) even with the ontological, or mechanistic, model. Such changes are well-known in chemical toxicology where it is referred to the "dose-dependent fate" of the toxic agent and is the result of changes with increasing dose (or dose/fraction) from 1) linear to non-linear, or Michaelis-Menten pharmacokinetics within a given metabolic route (linear $-dC/dt = k_0C$; non-linear, $-dC/dt = k_1C/(k_2+C)$, where C is the concentration and t the time) or 2) from one metabolic route to another with the onset of "saturation" effects: "With some chemicals, metabolic changes are induced by large doses rendering animals receiving such doses different from those receiving low doses. In either case, changes in the fate of the chemical or the metabolic status of the animal preclude the use of routine statistical processes to predict the hazard at low doses. Indeed, this violates the a priori assumption for their use" (Gehring et al, 1977).

In this context it is appropriate to note that we shall present in this report evidence that some of the published estimates of α/β that are obtained from the multifraction LQ model of radiation toxicity data are often "driven" by the observations at high levels of dose/fraction (D/N – 15 Gy) that are well beyond the stipulated range of validity of the LQ model ($0 \leq D/N \leq 10$ Gy) - an instance of extravagant extrapolation.

It is of importance - as well as of interest - to point out that the problem of extrapolation is simply one specific aspect of a fundamental problem of epistemology, namely, that of generalization, or induction, in which Hume (1745/1977) is the most celebrated, if not the first, to point out the absence of valid logical arguments that would justify the assertion that, "... those instances, of which we have had no experience, resemble those of which we have had experience," and, therefore, "... even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inferences concerning any object beyond those of which we have had experience." (Popper, 1965b). (N.B.: Recent studies in the newer fields of non-linear dynamics and non-equilibrium thermodynamics have disclosed the rich spectrum of recurrent behaviours (equilibrium, periodic, quasiperiodic, and chaotic motions) accessible via bifurcations (subtle, catastrophic, and explosive) that are available to non-linear dynamical systems maintained in far-from-equilibrium conditions, as a system parameter (e.g., a temperature gradient) is smoothly (and often only slightly) varied. The subtle bifurcations are analogous to second-order phase transitions and the catastrophic and explosive bifurcations are analogous to first-order phase transitions in equilibrium thermodynamics. Such studies dramatically reinforce the above remark by Box et al that "... the mechanism may change. Moreover, they also suggest the precocity of Hume's argument. See Thompson and Stewart, 1986 and Nicolis and Prigogine, 1989.)

3.2 Criteria for a scientific model

"The role of scientific methods is to restrain the investigators' advocacy." A. Feinstein, 1990.

It is also important that we have in mind a set of criteria for critically evaluating any model of dose-response. As can be inferred from the foregoing, parsimony is one important criterion. A useful, albeit quite general, set of criteria was described by Kuhn (1977) for evaluating a scientific theory. We have extended it to the evaluation of the models by which such theories are articulated and implemented: "What, I ask to begin with, are the characteristics of a good scientific theory? Among a number of quite usual answers I select five, not because they are exhaustive, but because they are individually important and collectively sufficiently varied to indicate what is at stake. First, a theory should be accurate: within its domain, that is, consequences deducible from a theory should be in demonstrated agreement with the results of existing experiments and observations. Second, a theory should be consistent, not only internally, or with itself, but also with other currently accepted theories applicable to related aspects of nature. Third, it should have broad scope: in particular, a theory's consequences should extend far beyond the particular observations, laws, or subtheories it was designed to explain. Fourth, and closely related, it should be simple, bringing order to phenomena that in its absence would be individually isolated and as a set, confused. Fifth - a somewhat less standard item, but one of special importance to actual scientific decisions - a theory should be fruitful of new research findings: it should, that is, disclose new phenomena or previously unnoted relationships among those already known. These five characteristics - accuracy, consistency, scope, simplicity, and fruitfulness - are all standard criteria for evaluating the adequacy of a theory." And, "The last criterion, fruitfulness, deserves more emphasis that it has yet received. A scientist choosing between two theories ordinarily knows that his decision will have a bearing on his subsequent research career. Of course he is especially attracted by a theory that promised the concrete successes for which scientists are ordinarily rewarded." (Kuhn, 1977). A repertoire of statistically adequate measures to implement each of these criteria are presented in Table 2. The respective statistical measures are fully discussed in parts 4 and 7 and Annexes II-IV of the present report.

Note that Kuhn's criterion of simplicity subsumes that of parsimony. Note as well, that for a useful predictive model the criterion of accuracy requires a closer "demonstrated agreement" of estimated and observed response than is implied by the failure of the data to reject the model at the usual level of significance ($\alpha = 0.05$) that obtains in most tests of hypothesis on goodness-of-fit of a model. Failure of the data to reject the model means only that the model is better than the mean response as a predictor of the response (vide infra).

Important Topics

Meta-analysis. Primary analysis. Secondary analysis. Epistemology. Ontology. LQ model. α/β ratio. Cross-over dose. Publication bias. Empirical models. Mechanistic models. Control models. Criteria for a scientific model. Generalized linear model. Linear predictor. Normal, Binomial, and Poisson distributions.

Table 2. Implementation of Model Criteria by Statistical Measures.

Criterion	Measure
1. Accuracy* (Concordance with sample data)	A) Sampling distributions of aggregate statistics. RSS = Σe_i^2 $\chi^2 = \Sigma \chi_i^2$ $D = \Sigma d_i^2$. $F = \Sigma g_i^2$ B) Plots of case statistics a) Plots of residuals, χ_i and e_i: i) $\chi_{(i)}$ vs z_i (Normal Probability plot) ii) χ_i vs X_j; e_i vs X_j iii) χ_i vs $\hat{m}_i$ or $\hat{\pi}_i$; e_i vs $\hat{y}_i$ iv) χ_i vs i; e_i vs i b) Plots of hat matrix diagonals, h_i i) h_i vs X_j ii) h_i vs χ_i; h_i vs e_i iii) h_i vs i c) Plots of components of parameter estimates $(\hat{\beta}_{(i)} - \hat{\beta})/\sqrt{Var(\hat{\beta})}$ vs i.
2. Consistency (Concordance with prior information)	A) Constraint: $\underline{r} = R\underline{\beta} + \underline{v}$. $E(\underline{v}) = \underline{0}$. $Var(\underline{v}) = \psi$. B) Posterior Odds Ratio: $P(M_1 \mid \underline{y})/P(M_0 \mid \underline{y}) = B*P(M_1)/P(M_0)$ where $P(\underline{y} \mid M_1)/P(\underline{y} \mid M_0) = B = L_1/L_0$. B is the Bayes Factor and L_1/L_0 is the Likelihood Ratio for the rival models M_0 and M_1.
3. Scope/Invariance (Concordance with new data)	$PRESS_1 = \Sigma e_i^2/(1-h_i)^2$ (Normal distribution) $PRESS_2 = \Sigma \chi_i^2/(1-h_i)$ (Binomial distribution) H_0: $\underline{\beta}(1) = \underline{\beta}(0)$ (Between-study invariance of parameter estimates)
4. Simplicity/Parsimony	$AIC_1 = -2\ln L - 2k$ $AIC_2 = \Sigma d_i^2 - 2k$ $AIC_3 = \Sigma \chi_i^2 - (n-k)$
5. Fruitfulness	$g(\mu_i) = \eta_i$**

* For models of non-Normal responses (Binomial, Poisson, etc.) the deviance, d_i, or Freeman -Tukey, g_i, residuals may, of course, replace the Pearson chi-squared residual, χ_i, in both aggregate and case statistics. χ^2, D, and F are each distributed asymptotically as Pearson chi-squared on (n-k) degrees of freedom where n is the sample size and k is the number of parameters β_j, $0 \le j \le k-1$, in the model.

[**] Rather than a statistical measure of the criterion of "Fruitfulness" we have given an ostensive measure in the generalized linear model (McCullagh and Nelder, 1983) which can be deployed to describe dose-response data in which the distribution of the random part, e_i, of the response, $y_i = \mu_i + e_i$, $1 \leq i \leq n$, may take a wide variety of distributional forms (Normal, Binomial, Poisson, etc.) and the deterministic part of the response, μ_i, may take a wide variety of linear (in β) forms. The response variable may be further generalized to describe data in which the treatment evokes responses in two or more systems of interest in the irradiated organisms; that is, in which the response of the irradiated organism must be represented by the vector (y_1, y_2) rather than the scalar, y.

N.B. The statistical measures are defined and discussed in sections 4 and 7 below.

3.3 Some principles of statistical modelling

"It is a mistake to suppose that the scientist somehow has a choice between statistical and non-statistical experiments."

E. A. Murphy, 1976

"Two phases of statistical modelling are model fitting and model checking. For linear models, model fitting is usually done via the maximum likelihood method and model checking includes diagnostics for departures from linear models (lack of fit), diagnostics for collinearity among columns of X, and diagnostics for model redundancy (model selection)."

C. Gu, 1992

We begin with an overview of some general principles for the practice of statistical modelling:

"Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. The first is that all models are wrong; some, though, are better than others and we can search for the better ones. [This is a re-statement of the well-known remark of G.E.P. Box: "All models are wrong but some are useful."] At the same time we must recognize that eternal truth is not with our grasp. The second principle ... is not to fall in love with one's model, to the exclusion of alternatives. Data will often point with almost equal emphasis to several possible models and it is important that the analyst accepts this. A third principle involves checking thoroughly the fit of the model to the data, for example by looking at residuals and other quantities derived from the fit to look for outlying observations, [emphasis added] and so on." (McCullagh and Nelder, 1983)

Still another, and most important, principle of modelling is the ancient and honorable "ontological tenet" of simplicity (Bunge, 1959), or parsimony - Ockham's Razor (14th century): "What can be done with fewer is done in vain with more." (It is of (some) interest to recall that William of Ockam (so named for the 14th century village in Surrey where he was born), the Franciscan friar and Late Medieval philosopher of science most widely remembered for his views on the principle of parsimony, was excommunicated in 1328 for his views on the vows of poverty.) Indeed, perhaps the strongest motivation to adequately summarize the n observations of a sample of data by the $k < n$ parameters of a statistical regression model is described by Mach's principle of economy of thought: It is easier to "think about" the few parameters of the model rather than the many observations of the sample in trying to give an exact and circumstantial account for the occurrence of the data in hand. It readily follows that in the selection of predictor, or explanatory, variables in the construction of a regression model, especially of non-experimental data, parsimony is a "traditional statistical principle" (Muirhead and Darby, 1987). However, the physicist Sir Harold Jeffreys (1960) appealed to the principle as an ontological tenet, his simplicity postulate ("The set of all possible forms of scientific laws is finite or enumerable, and their initial probabilities form the terms of a convergent series of sum 1 ... the order of decreasing initial probabilities is that of increasing complexity."), when he stated that the simplest of several rival mathematical forms for a natural law is the most likely to be correct ("... the simplest law is chosen because it is the most likely to give the correct predictions.") and, with Dorothy Wrinch (in 1921), defined the simplicity of a mathematical function by the paucity of its freely adjustable parameters - as do McCullagh and Nelder (1989) and everyone else now, also.

Especially in the construction of models of clinical data, in which both sample and non-sample information on the conditional response are weak, "... one may appeal to the principle of Occam's Razor, which states essentially that a model should be as 'simple' as possible; that is because lacking any other information a simple model has intuitively greater plausibility than a complex one" (Walter and Holford, 1978. See also Box, 1976). On the other hand, for the philosopher Sir Karl Popper, the simplest model is chosen because it is the most improbable - and hence most easily falsified. These two disparate views on the ontological role of simplicity (parsimony) are contrasted in the following schema:

1) Jeffreys-Wrinch
simplicity (of hypothesis) → paucity (of parameters) → high a priori probability
vs
2) Popper
low a priori probability → paucity (of parameters) → simplicity (of hypothesis)

Thus, historians, philosophers, physicists, epidemiologists, statisticians, etc., agree that, in modelling, it is usually the case that, "less is more", although the respective justifications thereof are 'sometimes (jarringly) different.

An empirical statistical argument for the Jeffreys-Wrinch interpretation of parsimony in modelling is, of course, that the model with fewer adjustable parameters captures less of the (conditional) random variation in the observed response (See Montgomery and Peck, 1982). Obviously, the more parsimonious model may also fail to capture some important information on response, as well, and well-informed judgment must be exercised in any application of the principle.

N.B. "But close concern with the data is needed as well: Occam's razor chooses the simplest among the explanations that fit" (T. Poston, 1979). The principle of parsimony has still more recently evolved into a formal model discrimination criterion, the Akaike Information Criterion (AIC) to be discussed below in section 7.7, in which both a measure of the goodness-of-fit of the model and data and the number of parameters included in the model are explicitly included.

There is, in any given application, the possibility that the principle of parsimony - as the tenet of simplicity - will be "over-read": Schlick (1961) has remarked that, "... any scientist who has succeeded in representing a series of observations by means of a very simple formula (e.g., by a linear, quadratic, or exponential function) is immediately convinced that he has discovered a law" - which is usually generalized, often uncritically, to other series of observations (the "illicit and hasty generalizations"?). There is some evidence of still other apparent "abuses" of the principle of parsimony. For instance, we may note that one of the characteristic weaknesses of current studies that was remarked above - too few subjects at risk in each arm of a clinical trial or at each level of an animal experiment - may be wryly (but usefully) viewed as due to a kind of excess of parsimony.

A fifth principle of modelling practice is to, "... choose a model that is consistent with the data and yields parameter estimates consistent with ... prior beliefs." [emphasis added] (J. Robins and S. Greenland, 1986). Or, as Leamer (1978) puts it: "It is apparent that a model is to be judged in terms of not only its R^2 [a measure of concordance of model and data] but also by the 'plausibility' [a measure of concordance of model and a priori information] of its estimates" [of the parameters and functions thereof]. This, to be sure, is an improvement on A. Osiander's (1540) recommendation: "Nor is it, to be sure, necessary that these hypotheses be true, or even probable; but this one thing suffices, namely, whether the calculations show agreement with the observations." (Andeas Osiander, in the preface to De Revolutionibus by Copernicus). Osiander's remark is one of the earliest expressions of the view of the so-called instrumentalist school (Popper, 1965a), which has dominated much of Western scientific thought (if not scientific practice, see T. Kuhn, 1970a) from the sixteenth through the twentieth centuries. (An instrumentalist holds that theories and laws have no truth in themselves; they are only instruments or intellectual artifacts for making calculations, organizing descriptions of experience, and for making inferences between past and future.) Instances of the instrumentalist view - as well as several of its transmogrifications - may be found in the current radiobiological literature - in those (rare) studies in which goodness-of-fit of the model and data is explicitly (though not always correctly) assessed and reported. A recent statement is Fowler's: "In any case, LQ is simply a low-dose approximation" (Fowler, 1989).

The statistician, John Tukey, has offered some useful comments on learning from data. They are worth repeating: "1. The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. 2) The data may not even contain the appearance of an answer, although we should look for appearances and then report them with adequate caution. 3) We must often expect to purvey appearances clearly labeled as such rather than answers." (J.W. Tukey, 1986).

4. Statistical methods. I. Parametric regression. The linear, non-linear, and generalized linear models.

'Statistics is ... primarily concerned with the theory and practice of matching theory to data."

J. Nelder, 1986

"The linear model has the elegant feature that each variable and parameter has a separate influence on the dependent variable."

W. Gilchrist, 1984

"Generalized linear models have unified the approach to regression for a wide variety of discrete, continuous, and censored response variables that can be assumed to be independent across experimental units."

S. Zeber and M. Karim, 1991

In order to present most concisely the conclusions reached in our assessment of the published findings of the received models we have examined it will be necessary to digress at several points in this report to introduce some statistical definitions and arguments, as well as methods and criteria, in our exposition of the matters at issue. The first of such digressions, which provides a discussion of a statistical model, follows. In order to be concise, the discussion is presented in matrix notation. Those readers for whom this notation is no longer familiar are referred to section 6.1.1 for a brief review.

In the modern metaphor, the observed response y_i in the i^{th} subject, $1 \leq i \leq n$, in a sample of size n, is the output of a process in which a deterministic signal, μ_i, is overlaid by random noise, ϵ_i. μ_i is the signal component of y_i that describes the level of response that is common to all subjects exposed to the treatment level, $\underline{x}_i^T$, and ϵ_i is the noise component that describes the level of the response that is unique to the i^{th} observation. μ_i and ϵ_i combine additively, $y_i = \mu_i + \epsilon_i$ (Or multiplicatively, $y_i = \mu_i * \epsilon_i$; however, we shall consider only those responses in which the signal and noise combine additively). Thus, the ϵ_i describe the deviation of the response of the i^{th} subject from the common (conditional) response μ_i; therefore $E(y_i) = \mu_i$ and $Var(y_i) = Var(\epsilon_i)$. The basic thrust of statistics is to extract the best representation of the signal, μ_i, from the data. "Success [in this enterprise] depends upon three components: (i) a good description of the signal, $[\mu_i]$, (ii) a good description of the noise, $[\epsilon_i]$, which together describe the model, and (iii) matching the model to the data." (Nelder, 1986)

Let us first describe the noise. Since the noise, ϵ_i, is a random variable, it is completely described by the form and parameters of its distribution. This is just the conditional distribution of y_i. For a sample of size n, the observed response y_i is a random variate with a conditional distribution of specified form in which the location parameter of the distribution is a function of one or more covariates $\underline{x}_i^T = (x_{1i}, x_{2i}, ..., x_{pi})$. The distributions of interest belong to the so-called exponential family and include the Normal distribution $N(\mu, \sigma^2)$, the Binomial distribution, $B(\pi, n)$ and the Poisson distribution, $P(\lambda)$. The parameters of the Normal distribution are μ (location) and σ (scale, $\sigma > 1$); the parameters of the Binomial distribution are π (location, Bernoulli probability, $0 \leq \pi \leq 1$) and n (Bernoulli trial); the parameter of the Poisson distribution is λ (location, $\lambda > 0$). The respective ranges of the response variates having these distributions are $(-\infty, +\infty)$ (Normal); (0[1]n)/n (Binomial); 0(1) ∞ (Poisson). The first two central moments of the Normal distribution are μ (mean) and σ^2 (variance). Similarly for the Binomial distribution we have $n\pi$ (mean) and $n\pi(1-\pi)$ (variance); for the Poisson distribution we have λ (mean) and λ(variance).

It is important to note that the deterministic part of the response, μ_i, is described by a function (linear or non-linear in the parameters) of the predictor variables, while the stochastic part, ϵ_i, is described by a statistical distribution function. Estimators of the parameters of the function of the predictor variables and of the parameters of the distribution function may be obtained from the sample or supplied by non-sample information.

The appropriate form of the distribution of ϵ_i is determined by the scale, or metric, on which the response is measured. If y_i is measured on a continuous scale $(-\infty, +\infty)$ then a Normal distribution is implied; if y_i is a proportion, then a Binomial distribution; if a count, but not in the form of a proportion, then a Poisson distribution, and so forth.

Let us now describe the form of the signal, that is, the form of the dependence of μ_i on the treatment variables $\underline{x}_i^T$, where $\underline{x}_i^T = (x_0, x_1, ..., x_p)_i$ is the (1*k) vector of treatment variables and other covariates, e.g., x_1 = D(Gy), x_2 = N(fraction), etc.

The deterministic, or structural, part μ_i of the response y_i describes how the response changes with changes in the treatment variables and covariates. The deterministic part of the response is a function, say $f(\underline{x}_i^T, \underline{\beta})$, of the (1*k) vector $\underline{x}_i^T$ and a (1*m) vector of unknown (free) parameters $\underline{\beta}^T = (\beta_0, \beta_1, ..., \beta_m)$, that is, $\mu_i = f(\underline{x}_i^T; \underline{\beta})$. The number, m, of parameters is much less than the number, n, of observations, m << n. Thus, the model achieves what the physicist Ernst Mach, has described as an economy of thought.

The function may be either linear or non-linear in the unknown parameter vector, $\underline{\beta}$. If $f(\underline{x}_i^T, \underline{\beta})$ is linear in the parameter vector, then m = k; if non-linear then $m \geq k$. Operational definitions of linear and non-linear regression models based on the derivatives of the expectation $E(\underline{y})$ with respect to the parameters $\underline{\beta}$ have been given by several investigators. We use those of Bates and Watts (1988). Linear Models: "... derivatives [of the expectation function] with respect to the parameters are independent of all the parameters." Non-linear models: "... at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters." (As a general remark, a linear model is usually regarded as a phenomenological approximation to the "true" mechanistic non-linear model of the process that generated the observations in the sample at hand.)

"Nonlinear regression models differ greatly in their estimation properties from linear regression models in that, given the usual assumption of an independent and identically distributed normal stochastic term, linear models give rise to unbiased, normally distributed, minimum variance estimators, whereas nonlinear regression models tend generally to do so only as the sample size becomes very large." Moreover, "Sometimes, the sample size required to achieve close-to-linear behaviour is far greater than would be practically possible." D. Ratkowsky, 1990. Indeed, for any given non-linear model it is impossible to know the size of sample required to achieve the required degree of "close-to-linear behaviour".

The non-linearity of a model has several sources, the two principal sources are the intrinsic non-linearity and the parameter effects non-linearity (See Bates and Watts, 1988; Ratkowsky, 1983, 1989). The latter can be altered by reparameterization of the model and this may often reduce the degree of non-linearity of the model - and hence of the bias in the estimates of the model parameters (vide infra). By reparameterization, it is meant that "... the parameters of one of them may be expressed as a function only of the parameters of the other model, without the expression containing the explanatory variables, the response variables, or the error term." (D. Ratkowsky, 1990). However, the predicted responses for the two parameterizations are the same. The issue thus, in part, becomes an issue of the relative merits of linear and non-linear models. As a general remark, non-linear models of data are to be avoided unless there is compelling theoretical support for the mechanism described therein. See for instance, Ratkowsky, "Non-linear Regression Modelling" (1983): "... in general the model that comes closest to behaving as a linear model will be the preferred choice." Ratkowsky notes further that non-linear regression models differ in general from linear regression models in that the, "... estimators of the parameters are biased, non-normally distributed and have variances exceeding the minimum possible variance." The, "... instability in parameter estimates is closely related to non-linear behaviour of the model." Ratkowsky (1983) has shown that the bias in the sample estimates of the parameters of the reparameterization is a strong function of the bias and variance of the rival parameterization. In fact, Ratkowsky (1983) has shown that the bias in the estimates of the parameters of a non-linear model may exceed 100% of the estimate! The received method for obtaining reduced-bias estimators of the parameter vector of non-linear regression models is the Quenouille version of the jackknife. See Bard, "Non-linear Parameter Estimation", 1974. See Also Hinkley, 1977. N.B.: There is a kind of Uncertainty Relation that obtains for non-linear models: Although it is acknowledged by most investigators that non-linear models are more likely to be correct - since "Nature is non-linear" - the degree of bias, variance, and non-Normality of the distributions, of the sample estimates of the parameters of such models are likely to greatly exceed those of the sample estimates of the cognate linear models - for sample sizes that are likely to be obtainable in most investigations.

However, although we shall provide some examples of the construction, testing, and deployment of non-linear models of isoeffect data when we consider the so-called linear-quadratic, or LQ, hypothesis, for most of the report we shall be more interested in the so-called generalized linear models of dose-response data in which $g(\mu_i) = \eta_i = \underline{x}_i^T\underline{\beta}$ where g(.) is a link function for which the form is determined by the nature of the distribution of ϵ_i. If the ϵ_i have a Normal distribution, then g(.) is the identity function, $g(\mu_i) = \mu_i$. If the ϵ_i have a Poisson distribution then $g(\mu_i) = \log \mu_i$. If the ϵ_i have a Binomial distribution then $g(\mu_i) = z_i$, the probit transformation, $z_i = \Phi^{-1}(\pi_i)$, where $0 \leq \eta_i \leq 1$ and $\Phi(.)$ is the Normal distribution function. (The logit transformation, $z_i = \log[\pi_i/(1-\pi_i)]$, is also appropriate). π_i is the so-called linear predictor. It is a linear function of the parameter, $\underline{\beta}$. Thus, the generalized linear models are defined by three properties: 1) the error structure, that is, the form of the distribution of ϵ_i; 2) the linear predictor, π_i; 3) the link function, g(.). As an example, consider the case where the response, y_i, is the number of successes, r_i, out of n_i independent trials, where π_i is the probability of a success in a single trial, that is, $y_i - B(n_i, \pi_i)$, or, $y_i = n_i\pi_i + \epsilon_i$, where ϵ_i represents binomial deviations from the expectation $E(y_i) = n_i\pi_i$. If π_i is conditional on a dose x_i then we may write $\underline{y}_i/n_i = \Phi(z_i) + \epsilon_i$ where $z_i = (x_i - \mu)/\sigma = \eta_i = (1/\sigma)x_i - (\mu/\sigma)_T = \beta_0 + \beta_1 x_i = \underline{x}_i^T\underline{\beta}$. Here (μ, σ) are the parameters of the Normal tolerance distribution and $\underline{x}_i^T = (1, x_i)$.

Epistemologically, the generalized linear models are, in the main, phenomenological models. The equation, $g(\mu_i) = \eta_i$, expresses a direct connection between the observed quantities, or their respective metameters, or transforms. (The received version of the multifraction, $N \geq 1$, LQ model is a mechanistic model since it postulates underlying occurrences which are relatively simple to visualize - but not necessarily to observe: $S_N = \Pi S_1$ where S_1 is the survival at one dose fraction and S_N is the survival at N dose fractions. Indeed, the response, $-\ln S_N$, in the received multifraction LQ model (vide infra) often refers to an unobservable level of survival S_N in an indefinable cell population.)

With these definitions in hand we can now specify one of the criteria - accuracy - for an adequate model more concisely: We have an adequate model if a plausible parameter vector $\underline{\beta}$ exists such that the model $f(\underline{x}_i^T, \underline{\beta})$, $1 \leq i \leq n$, maps the observations into a white noise sequence unrelated to any known variable. In symbols we have:

$$\underline{\beta} : (y_i, \underline{x}_i^T) \longrightarrow \epsilon_i.\ 1 \leq i \leq n.$$

Or, in matrix notation:

$$\underline{\beta} : [\underline{y}, X] \longrightarrow \underline{\epsilon}$$

where $[\underline{y}, X]$ is the n*(k+1) observation matrix, $\underline{y}$ is the (n*1) response vector, X is the (n*k) matrix of predictor variables, and $\underline{\epsilon}$ is an (n*1) noise vector. By "plausible" we mean that $\underline{\beta}$ is consistent with a priori information on the process. (It will be recalled that accuracy and consistency are the first two desiderata of a good scientific theory that were cited by Kuhn.)

A model provides an account of the observed variation in the response over the range of observations. The deterministic part describes a "law" that governs that variation, and, "The crucial point is that ... the greater part of the variation is accounted for by the law, leaving an unexplained balance. The ground for accepting and generalizing the law is therefore a quantitative one: how much of the observed variation needs to be explained by the law before we can accept the law? The differences are called errors (in no derogatory sense) or residuals, and before we can proceed any further we must consider their properties." (Jeffreys, 1957). Perhaps the most famous residual in the physical sciences is the error of eight minutes of arc between the observations of Tycho Brahe and the estimates based on the Ptolemaic hypothesis: "... these eight minutes alone have prepared the way for an entire reform of Astronomy and are to be the main subject of this work." (J. Kepler, De Stella Motibus, 1609).

The residuals are the statistics that can disclose the degree to which the model matches the theory to the data, or, in Osiander's locution, "... whether the calculations show agreement with the observations" (vide infra). For parametric models of dose-response, the i^{th} residual, say e_i, $1 \leq i \leq n$, is defined as the difference between the observed level of response, say y_i, and that predicted by the model, say $\hat{y}_i = \hat{\mu}_i = \underline{x}_i^T\hat{\beta}$, at $\underline{x}_i^T$: $e_i = y_i - \hat{y}_i = y_i - \hat{\mu}_i$. Here $\hat{\beta}$ denotes the sample estimate of β (to be obtained by least squares or maximum likelihood methods as described in section 7). (For parametric models of time-to-failure, the i^{th} residual is defined as the ratio $y_i/\hat{\mu}_i = e_i$, Cox and Snell, 1968.) For $e_i = y_i - \hat{\mu}$, the sum of squared residuals, RSS = Σe_i^2, provides an aggregate statistical measure of goodness-of-fit of model and data whose interpretation depends upon the form of the distribution of the ϵ_i.

Sample estimates, $\hat{\beta}$, of the parameter vector, β, are obtained by mathematical methods that select values for β that either maximize or minimize a given criterion function, i.e., that either minimize the sum of squared residuals, Σe_i^2, or maximize the likelihood of occurrence of the sample. For data in which the random part of the response has a Normal distribution, i.e., the linear and non-linear regression models, the parameter estimates are obtained by least squares methods. For Binomial and Poisson responses, i.e., the generalized linear models, maximum likelihood methods are used. The former are single-step methods; the latter are iterative. Estimation methods are discussed at greater length in section 7.

The set of residuals $[e_i]$, where $e_i = y_i - \hat{\mu}_i$, $1 \leq i \leq n$, is the i^{th} residual, represent all of the information in the data that is not included in the parameter estimate $\hat{\beta}$. That is, the $[e_i]$ "... contains all the available information on the ways in which the fitted model fails to properly explain the observed variation in the dependent variable y_i" (N. Draper and H. Smith, 1981). The residuals e_i, $1 \leq i \leq n$, are case statistics since there is one statistic for each of the n cases, or observations in the sample. Case statistics are also referred to as regression diagnostics.

Analysis of the residuals seeks to detect and identify the misspecifications of both the deterministic and random parts of the response $y_i = \mu_i + e_i$. With respect to the former it seeks first to detect the misclassification of deterministic components, say x_j, as stochastic components and then "migrate" them from ϵ_i to μ_i. Symbolically,

$$\begin{tikzcd}[column sep=small]
y_i = & \mu_i & + & \epsilon_i \arrow[dll, bend left] \\
& x_j \arrow[u, bend left, "?"'] & &
\end{tikzcd}$$

Such misspecification usually is detected and identified by the presence of "pattern" in plots of the set of residuals, such as e_i vs $\hat{\mu}_i$, and e_i vs x_j, $1 \leq i \leq n$, where $\hat{\mu}_i$ is the estimate of μ_i given by the model. If there is no pattern then it is assumed that the model has captured all of the information present in the data - since none has escaped into the residuals. There is, of course, the problem of "over-reading" the set of residuals: If one looks hard enough some pattern can be discerned in any finite set of residuals - just as one can see "faces" in clouds.

Probability plots of e_i are useful in detecting misspecification of the form of the distribution of the random part of y_i, $1 \leq i \leq n$. These are plots of the cumulative distribution of the residuals, the ordered e_i vs $i/(n+1)$, on a special grid having the property that if the distribution of e_i is the same as that on which the rulings of the grid are constructed, say the Normal distribution, then the plot of e_i vs $i/(n+1)$ is a straight line. In probability plots, the presence/absence of a linear pattern in the plot is the criterion on which the form of the distribution of the e_i is assessed. We shall discuss residual analysis at greater length in section 7. Examples of the use of residual plots are given in sections 7-10 (see, for example, figs. 5 and 6 and 19 and 20).

5. "Why do we believe it?"

"Of course I can believe a thing without understanding it. It's all a matter of training."

D. Sayer, 1936

"It is not what the man of science believes that distinguishes him, but how and why he believes it."

Bertrand Russell, 1930

The secondary analyses of those studies listed in Table 1 are described in Annexes II-IV. They have disclosed that, on normative measures and criteria, neither the theoretical arguments nor the empirical evidence presented in most - but not, of course, all - of those reports can support the respective (published) conclusions; that is, the conclusions and the evidence often fail to intersect. Indeed, in not a few of the studies the evidence is logically - and obviously - devastating to the conclusions that are based on it. Therefore, for those studies, it would seem that we are obliged to account for the striking fact that although on normative criteria, such as analysis of the several plots of the residuals, say e_j vs $\hat{\mu}_j$, and the sum of squares Σe_j^2, the conclusions to those studies are "unbelievable", they are, nonetheless, firmly believed and widely deployed by distinguished investigators, anyway - since they are published in peer-reviewed journals, and frequently cited in authoritative reviews.

It is useful to restate our perceived dilemma in terms of the views of a theoretical physicist, J. Ziman, FRS, on the nature of Science. First, the fact that the studies listed in Table 1 are published had unambiguously identified them with that consensus that Ziman (as well as others) have identified as distinguishing the practice of science from all other artistic and intellectual pursuits and pastimes: 1) "Science is unique in striving for and insisting on a consensus" (J. Ziman, 1968). 2) "... a fundamental principle of science is general acceptance." (H. Jeffreys, 1957). 3) "Scientific knowledge, like language, is intrinsically the common property of a group or else nothing at all" (T. Kuhn, 1970a). Consensus functions as the "bottom-line" criterion for the scientific enterprise.

However, Ziman (as well as others) have argued that the uniquely persuasive rhetorical power of the scientific method - in achieving consensus - lies in the demonstration of agreement between observation and prediction: "Scientists continually refer to 'agreement between theory and experiment'. This is what makes a scientific paper convincing: a pattern of theory is shown to imply features which are confirmed by experiment" (J. Ziman, 1968)[3].

But for many of the studies that are currently published the required "agreement between theory and experiment" is not - and for many cannot be - demonstrated on statistically adequate criteria. Thus, the insistent question arises, What other - non-normative - persuasive rhetorical powers are available to investigators seeking to achieve the required consensus? Some answers to this question are given below.

It should be remarked at once that it appears to us that for the models at issue in this report, the striving by the peer-group to achieve consensus has prematurely stifled debate on several important issues in radiation dose-response; it is the case, of course, that "... acceptance of an inadequate explanation discourages search for a good one." (H. Jeffreys, 1957).

It is also the case, that philosophers and historians of science have long devoted much well-informed conjecture and speculation - as well as weighty argument - to attempts to provide a sufficient answer to the insistent question of why one scientific theory - but not another - was believed by the peer-group of a given field or specialty at a given epoch in its development. Their studies have disclosed that there are at work other cognitive and emotional processings of one's experience that can give rise to those "feelings of conviction" that describe the subjective experience of belief that are quite distinct from the unprejudiced observations advocated by Bacon.

It will suffice for the purposes of the present report to briefly describe several of those "para-normative" processings in order to resolve the above dilemma and account for those instances of "believing the unbelievable" which we believe that we have demonstrated in Appendix I and Annexes II-IV. Thus, we offer a summary of some quite early, as well as some more recent, accounts of the formation and maintenance of scientific beliefs, together with some evidence that suggests that these views may provide a sufficient answer to our second question: why the received models are so. However, we must stress the conjectural and heuristic nature of our answer.

5.1 Some views on the formation and maintenance of beliefs

We present first several of the views of the aristocratic solicitor whom most consider to have co-founded - with the gentleman-soldier René Descartes - the "going concern" that is Western, that is, empirical, science: Baron Verulam, Francis Bacon. His views appear to us to still have, in addition to an engaging grandiloquence, considerable explanatory power in the matter at issue, namely, the para-normative foundations of belief. The views are taken from his Novum Organum, first published in 1620:

"45. The human understanding, from its peculiar nature, easily supposes a greater degree of order and equality in things than it really finds; and although many things in nature be sui generis and most irregular, will yet invent parallels and conjugates and relatives, where no such thing is. ..."

"46. The human understanding, when any proposition has been once laid down (either from general admission and belief, or from the pleasure it affords), forces everything else to add fresh support and confirmation; and although most cogent and abundant instances may exist to the contrary, yet either does not observe or despises them, or gets rid of and rejects them by some distinction, with violent and injurious prejudice, rather than sacrifice the authority of its first conclusions. ... Besides, even in the absence of that eagerness and want of thought (which we have mentioned), it is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives, whereas it ought duly and regularly to be impartial; nay, in establishing any true axiom the negative instance is the most powerful."

"47. The human understanding is most excited by that which strikes and enters the mind at once and suddenly, and by which the imagination is immediately filled and inflated. It then begins almost imperceptibly to conceive and suppose that everything is similar to the few objects which have taken possession of the mind, whilst it is very slow and unfit for the transition to the remote and heterogenous instances by which axioms are tried as by fire, unless the office be imposed upon it by severe regulations and a powerful authority."

Bacon's earlier philosophical conjectures and excursions have been refined - and reinforced - by the more recent and more solid scientific work in cognitive psychology by Nisbett and Ross (1980), and others, on belief and the epiphenomenon of belief perseverance:

"People tend to persevere in their beliefs well beyond the point at which logical and evidential considerations can sustain them.

1. When people encounter probative evidence pertinent to prior beliefs they tend to apply asymmetric critical standards to supportive and opposing evidence and tend to become more confident of a belief in response to a set of mixed evidence which normatively should serve to lower confidence.

2. People do not observe the 'commutativity' rule in response to sequentially presented evidence. Instead, early-presented evidence seems to create theories which are not revised sufficiently in response to later-presented, conflicting evidence.

3. Beliefs tend to sustain themselves even despite the total discrediting of the evidence that produced the beliefs initially.

Belief perseverance sometimes seems to occur because people have an emotional commitment to the belief. Perseverance is likely even when there is no such investment, however, because (a) people tend to seek out, recall, and interpret evidence in a manner that sustains beliefs, (b) they readily invent causal explanations of initial evidence in which they then place too much confidence, and (c) they act upon their beliefs in a way that makes them self-conforming."

With respect to 3(b) Nisbett and Ross (1980) note that, "People's facility in forming causal explanations is so great that they usually will be able to explain most events and relationships they observe. These explanations may often prove so convincing that they survive even the total discrediting of the 'evidence' that prompted their invention in the first place." Or, as Tversky and Kahneman (1982) remark, "It is a psychological commonplace that people strive to achieve a coherent interpretation of the events that surround them, and that the organization of events by schemas of cause-effect relations serves to achieve this goal. ... classic work ... provided a compelling demonstration of the irresistible tendency to perceive sequences of events in terms of causal relations, even when the perceiver is fully aware causality is illusory."

5.1.1 Null hypotheses

We shall have occasion to refer to "the null hypothesis" at several places in the present report; therefore we interpolate a brief discussion of the concept at this point. "In statistical hypothesis testing, the null hypothesis most often refers to the hypothesis of no difference between treatment effects or of no association between variables" (Greenwald, 1975). If the data reject the null hypothesis, then the investigator may conclude, with a specified probability of a Type I error (the level of significance of the test), that there is a difference between the treatment effects or that there is an association between the variables (N.B. Rejection of the null hypothesis signifies only that the difference or the association is "real" - not due to chance - not that either are of a size that may be "useful" to the investigators' sponsors.) On the other hand, in tests of goodness-of-fit of a regression model to a sample of data, the null hypothesis is the hypothesis that the model is correct - that it "fits" the data. In this case, if the data reject the null hypothesis the investigator may conclude, again with a specified probability of a Type I error, that the model and the sample refer to different populations. ("To test whether the line is an adequate representation of the data, a χ^2 text may be used, ... A value of χ^2 within the limits of random variation indicates satisfactory agreement between theory (the line) and observation (the data)" (Finney, 1971/1980).) Note that "within the limits of random variation" implies that the probability, $P(\chi^2)$, lies within the interval $0.05 \leq P(\chi^2) \leq 0.95$; i.e., the agreement between theory and observation may be so close that the data reject the model. See below. Note also that failure of the data to reject the null hypothesis signifies only that the model is a better predictor of the observed response than is the mean of the sample responses - not that the model is necessarily a "useful" predictor. See discussion of the Box-Wetz (1973) predictive criterion below.)

There is abundant evidence in the literature that there is, among both investigators and journal editors, a kind of exuberant "epistemological optimism" that finds expression in a marked prejudice against the null hypothesis as defined by Greenwald and for the null hypothesis as defined by Finney. That is, there is a prejudice to report and to publish the "good news" that the treatment effects are different, the variables are associated, the model is valid. Therefore, there is a prejudice in favor of "positive" rather than "negative" results - as remarked by Bacon (1620). Holton (1978) notes also that, "On looking back at her [Anne Roe] long and distinguished studies on the psychology of scientists, she is said to have commented that the one thing that all of these very different people had in common was an unreasonable amount of optimism concerning the ultimately successful outcome of their research."

Thus, it may often be the case that only a little training is required in order to, "believe a thing without understanding it". Well-developed innate capacities exist for inventing a) cause-and-effect relationships and b) "... parallels and conjugates and relatives, where no such thing is" (capacities which have had no little survival value for those possessing them to moderate degrees), coupled to a lusty appetite for positive findings ("good news") will often do the jobs required for the formation and maintenance - "perseverance" - of a stable belief system - often even despite strong theoretical and empirical evidence to the contrary.

At this point it is of more than passing interest, perhaps, to emphasize that 1) the epistemologies of both Bacon and Descartes[4] which provide the under-pinnings of Western science are, in outlook, basically optimistic - although in any given instance the optimism is usually misplaced (Popper, 'Sources of Knowledge and Ignorance," 1965); 2) Peirce notes ("The Fixation of Belief", 1877) that, as a general remark, most of the rest of us are wont to take too much counsel from Hope: "Most of us, for example, are naturally more sanguine and hopeful than logic would justify. We seem to be so constituted that in the absence of any facts to go upon we are happy and self-satisfied; so that the effect of experience is continually to correct our hopes and aspirations. Yet a lifetime of the application of this corrective does not usually eradicate our sanguine disposition. Where hope is unchecked by any experience, it is likely that our optimism is extravagant."; and that 3) the term gospel, meaning something accepted as unquestionably true, derives from the Old English "good news". Moreover, a recent study by Couch et al (1981) on surgical misadventures notes that, "The identification of the origins of error is not susceptible to scientific analysis,but throughout this enquiry many such factors were apparent and recurrent ... First of all, there is misplaced optimism ..."

5.1.2 Some evidence for the relevance of these views on belief in the present context

We include here, a few examples, taken from those papers and reviews listed in Table 1 in which it is evident, to even a cursory examination, that either the study data, or the statistics derived therefrom, that are presented in a published paper as evidence for the conclusions presented therein, do either gravely weaken, or even confute altogether, the published - and invariably positive - conclusions of the investigator(s) as to what these data mean in the context of the matter at issue in the study. In other words, the empirical evidence presented would, normatively, require that the published conclusions (estimates, inferences, etc.) either be accepted only with strong reservations or rejected outright. However, the published conclusions are positive and/or the preferred model is simple. These are the emotive features of the conclusion that completely overwhelm the cognate information in the data and/or statistics that are presented and, on normative criteria, either weaken or negate the conclusions.

a) In the Tucker and Thames (1983) paper on radiation toxicity (see Figs. 9 and 10 below) it is obvious that: i) The assertion that, if the respective F_e-plot is "straight", then the multifraction, $N \geq 1$, LQ model is a valid representation of the dose-response information in the sample is a logical fallacy - "affirming the consequent". ii) At only 7 of 19 treatment regimens, $\underline{x}_i^T$, $1 \leq i \leq n = 19$ do we find that $0 < r_i < n_i$, that is, an excess of the data resides in extreme responses: $r_i = 0$ or $r_i = n_i$. iii) At 9 of the 19 regimens, the dose per fraction, D/N, exceeds 10 Gy, the level of D/N that is stipulated as the upper limit of the range of validity of the multifraction LQ model (Fowler, 1984; 1989). iv) The ratio α/β is unique only to within a multiplicative constant - and therefore may not be a good discriminator for any purpose. v) The distribution of levels of dose, D, within each of the seven levels of (N, T) that comprise the experiment is quite bizarre since some levels of (N, T) have as few as 2 levels of D and one has as many as 4. vi) The variables (N, T) are highly correlated; therefore, it will be difficult to disentangle their separate roles in modulating the effects of dose on response.

b) In the Thames et al (1982) paper it is apparent that, again, the sample estimate of α/β obtained from the F_e-plot (see Fig. 22 below) is dominated by the observation at $D/N \simeq 15$ Gy, well above the stipulated range of validity of the multifraction LQ model ($0 \leq D/N \leq 10$ Gy). It is obvious from the F_e-plot that deletion of the observation at $D/N \simeq 15$ Gy will materially change the estimate of α/β (in the event, by a factor of – 2).

c) It is evident from Table V-8 of the BEIR III (1980) report that the model of choice, LQ-L, overfits the leukemia incidence data since the difference in the respective chi-squared statistics, $\chi^2(\text{df})$, for the LQ-L model and either the L-L or Q-L models is less than 3.84, the 0.95 quantile of the distribution of $\chi^2(1)$ - chi-squared on 1 df (see also Fig. 21a below).

d) It is evident from Fig. 5 of the Shellabarger et al (1969) paper that the dose-response curve for mammary neoplasia is linear, as is maintained in NCRP 64 (1980), only if the observations at the two extremes of dose are ignored (see Figs. 13a and 30a below). Indeed, to state that the dose-response curve for the incidence of mammary neoplasia in female Sprague-Dawley rats exposed to Co^{60} radiation is linear over the range 16R-250R is, to borrow Churchill's locution, "true but not exhaustive". It is a highly conditional truth; the condition being that the observations at the two extremes of dose have been deleted from the sample from which the linear model was constructed. Moreover, the linearity hypothesis is still quite dubious even when so circumscribed by this condition. The set of doses in the remaining observations are $\underline{D}^T = (15.625, 31.25, 62.5, 125, 250)$ with successive ratios (2, 2, 2, 2). The cognate set of responses is $\underline{\pi}^T = (0.06, 0.06, 0.17, 0.33, 0.59)$ with successive ratios (1.00, 2.83, 1.94, 1.79). However, the Shellabarger et al (1969) paper asserts that, "As the dose is doubled the response doubles approximately." Obviously, at the lowest doses, the region of greatest interest, doubling the level of dose (from 16 to 32) does not change the level of response! Or, as remarked by Nisbett and Ross (1980), "People tend to persevere in their beliefs well beyond the point at which logical and evidential considerations can sustain them."

It is of interest that at least one very distinguished scientist has offered a causal explanation of the putative linearity of the radiation dose-response curve for mammary neoplasia in the female Sprague-Dawley rat: "... this linear dose response could be due to the endogenous promotion by hormones in breast tissue" (Kennedy, 1985). Or, as remarked by Nisbett and Ross (1980), "People's facility in forming causal explanations is so great that they usually will be able to explain most events and relationships they observe. These explanations may often prove so convincing that they survive even the total discrediting of the 'evidence' that prompted their invention in the first place." It is of some interest, in the present context, to observe that Woolf (1988) has remarked that, "Copy editors report that they often notice data that appear to be 'too good' or graphical or tabular material that disagrees with the text." (emphasis added)

As a general observation, statistically adequate evidence that any model "fits" the data cited in the study as warrant for the reported results is rarely ever presented in any published report! The "agreement" is assumed rather than demonstrated by most investigators. If "consensus" - the object of the exercise - can be achieved without such demonstrations, there is, of course, no need for them; statistically adequate measures of "fit" are supererogatory. On the other hand this received practice might be justified on less secular arguments that suggest that the peer-group takes an excessively Cartesian view of their work, a view aptly summarized by Peirce (1877/1955) as follows: "In particular God cannot be a deceiver; whence it follows that whatever we quite clearly and distinctly think to be true about any subject must be true."[4] (Descarte's views were transmogrified into Hegel's so-called philosophy of the identity of reason and reality: "That which is reasonable is real, and that which is real is reasonable; thus reason and reality are identical." See Popper, 1970. There is an unmistakable Hegelian tinge to several of the reports listed in Table 1. See also Hibben, 1984.)

5.2 Roles for "positive findings" and "simplicity"?

"Seek simplicity - and distrust it."

A. N. Whitehead, 1927

"For every complex question there is a simple answer - and it is wrong."

H. L. Mencken, 1930

Our assessment, as well as that of others, strongly suggests that investigators often need only report a positive effect in order for their findings to be believed; the "positivity" being a sufficient, as well as a necessary, warrant for their "believability" - no other warrants are required. In fact, other warrants that refute believability are often ignored, as remarked above in section 5.1.1 and instanced in 5.1.2.

However, as has been pointed out by many since Bacon, "... in establishing any true axiom the negative instance is the most powerful." (N.B. The "negative instance" refers to the modus tollens of classical logic.) Indeed, the central thesis of Sir Karl Popper's Logic of Scientific Discovery rests on "... the asymmetry of a generalization and its negation in their relation to empirical evidence. A scientific theory cannot be shown to apply successfully to all its possible instances, but it can be shown to be unsuccessful in particular applications" (Kuhn, 1970b). That is, universal statements (hypotheses or theories) "... are never derivable from singular statements [accounts of the results of observations or experiments] but can be contradicted by singular statements. Consequently it is possible by means of purely deductive inferences (with the help of the modus tollens of classical logic) to argue from the truth of singular statements to the falsity of universal statements. Such an argument to the falsity of universal statements is the only strictly deductive kind of inference that proceeds, as it were, in the 'inductive direction'; that is, from singular to universal statements" (Popper, 1965a). This suggests that in these cases the formation and maintenance of the system of beliefs may contain one or more logical - intellectual - errors.

Greenwald (1975) presents, in paraphrase, some of the reasons why the prejudice against the null hypothesis - and in favor of "positive findings" - is so widespread. Although he subsequently shows each of them to be in error, they are worth reproducing here since they provide another facet of the answer to our question of, "Why do we believe it?" In the context of his remarks the null hypothesis is that the effects of two treatments do not differ, the two variables are not associated, etc.:

"1. Given the characteristics of statistical analysis procedures a null result is only a basis for uncertainty. Conclusions about relationships among variables should be based only on rejections of null hypotheses."

"2. Little knowledge is achieved by finding out that two variables are unrelated. Science advances, rather, by discovering relationships between variables."

"3. If statistically significant effects are obtained in an experiment, it is fairly certain that the experiment was done properly."

"4. On the other hand, it is inadvisable to place confidence in results that support a null hypothesis because there are too many ways (including incompetence of the researcher), other than the null hypothesis being true, for obtaining a null result."

Given the well-documented ubiquity of the prejudice against the null hypothesis, many investigators are concerned that perhaps much of the scientific literature on some important issues may consist of Type I errors. That is, studies for which the positive result that was achieved - and reported - was due to chance whereas other studies of the same issue for which a negative result was achieved were not published but remain in the investigators files owing to the fact that, as remarked above, editors as well as investigators are, "... more moved and excited by affirmatives than negatives, ..." In a "worst-case" view of the problem, the so-called "file-drawer problem", the journals are filled with the 5% of the studies that show Type I errors while the file drawers are filled with the 95% of the studies that show "non-significant" results (Rosenthal, 1979). Quantitative procedures for computing the stability, or tolerance, of the empirical evidence for a positive effect for filed and future non-significant studies have been reported by Rosenthal (1979) and by Begg (1985). "Unfortunately, these attempts to correct for publication bias can only provide a rough guide since they are based on assumptions about the missing data" (Simes, 1987).

However, one conclusion from the Rosenthal (1979) study that is important for the issue of believability in any concept is the following: Small numbers of studies that achieve positive results that are not very significant may well be misleading in that only a few negative studies that were filed rather than published could change the combined significant result of the published studies into a non-significant one.

As well as the "positivity" of study findings in general, there is also some evidence to suggest that, per Schlick, in studies on the construction of dose-response models, the "simplicity" of the model at issue is widely accepted as a warrant for its "believability". The principal evidence for this conclusion is that, as remarked in 5.1.2e above, the concordance of the model with any data is rarely evaluated by any criterion whatever and still less frequently evaluated by any statistically adequate criterion. We interpret this omission as some evidence of the investigator's sincere belief that the dose-response models which they deploy are, simply because they are simple, also "law-like".

Now, it is, of course, the case that simplicity is indeed one of the more widely accepted criteria for the validity of a model (Kuhn, 1977). Jeffreys himself has put its case quite forcefully: "I say on the contrary, that the simplest law is chosen because it is the most likely to give correct prediction; that the choice is based on a reasonable degree of belief ..." (Jeffreys, 1961). However, the choice of a model is never made on the criterion of simplicity alone - as Jeffreys notes (and Whitehead warns against).

5.3 Intellectual vs motivational errors

"Numberless, in short, are the ways, and sometimes imperceptible in which the affections color and infect the understanding."

F. Bacon, 1620

But,

"Who can understand his errors .."

Psalms 29: 12

Despite the markedly aleatory, or contingent, nature of the outcome to many biological processes - and, in particular, of the "outcome" in any dose-response process as remarked above - one must conclude from their reports that most investigators who are currently engaged in the construction, evaluation, and deployment of models of dose-response are possessed of only a weakly *stochastic* world-view. Cogent evidence for this conclusion is found in the features that we have noted earlier: the exiguous numbers of observations that furnish the empirical bases of most published studies - both clinical and animal - and in the often conspicuous absence of any measure whatever of the inherent ambiguities which must, perforce, encumber the choices of 1) the dose-response model for a given set of data, 2) the estimates and inferences derived therefrom, and 3) the a priori information on conditional response, that are all regularly reported - with apodictic assurance - in the professional literature. That is, most investigators fail either to obtain a sufficient number of observations to circumscribe the effects of chance on their findings, or to adequately describe the extent to which they do. For example, it is most unusual to find in the reports of such studies the standard statistical measures, such as *goodness-of-fit* (of model and data), *confidence limits* (on estimates of parameters), etc. Similarly, although *a priori* information on the parameters of a regression model is fairly often deployed to stabilize - or otherwise to "improve" - the sample estimates of the parameters, it is even more unusual to find in published reports, any measures of either the *uncertainty* in the prior information or of its *consistency* with the cognate sample information. (It should also be remarked that, although such a priori information is often so deployed, there is never any explicit acknowledgement of their maneuver by the investigators, suggesting that, "They know not what they do.") Their evident lack of concern for the stochastic ambiguities may well devolve from the fact that *many of the better-known investigators in this field were professionally trained in the more deterministic disciplines*, e.g., in physics, in chemistry, or in engineering (Sir Harold Jeffreys, a physicist of no small renown, has observed of many of his colleagues that, "It must be remarked that physicists hardly ever give a statement of uncertainty derived from the observations in the way required. If they give any it is usually little more than a guess, and quite useless for a significance test. A proper estimate of uncertainty can be recovered from the original readings, if these are published, but more often they are not. A physicist will cheerfully spend months on an experiment and grudge the few hours needed to present his results in the form that would make them of full use to other people." And another distinguished physicist, P.M.S. Blackett, once said, with only slight exaggeration intended, that, "... a physicist is satisfied with an argument if it leads to a result that he believes to be true" Kotz and Johnson, 1982). It is unfortunately the case that, for most investigators trained in the physical sciences, statistical methods are often dismissed as, "... merely the defective but necessary tools of people who know too little." (Hacking, 1990) with the implication that such people were the secondary workers in science who could never "know" more.

It is important to distinguish *intellectually*-based errors from *motivationally*-based error, since they are differently corrigible although their effects on the bias and precision of the reported estimates and inferences obtained from regression models of dose-response data may be indistinguishable. For example, an investigator who presents an *incorrect measure* of the goodness-of-fit of a model and data commits an *intellectual error*. An investigator who fails to present *any measure* of goodness-of-fit, commits a *motivational* error - at the very least he suffers from a volitional defect. But in either case, the investigator's report fails to answer the all-important question, Does the model "fit" the data?

Two other, less precedented, examples, which will be discussed at greater length in subsequent sections of this report, will be briefly noted here. 1) An investigator who unwittingly deploys a "fitting procedure" that arbitrarily *inflates* the weight assigned to a given observation in his sample commits an *intellectual* error (See Fig. 12b). 2) An investigator who arbitrarily deflates the weight to be assigned to one or more observations (in order to improve the "fit" of his selected model) commits a motivational error (See Fig. 13a).

Intellectual errors can be corrected by further instruction. The correction of motivational errors requires persuasion and incentives to achieve a change in goals. This task group report provides both instruction and incentive. Instruction in the appropriate statistical concepts, methods, and criteria is provided in sections 7-9. Incentive is provided by the evidence of widespread occurrence of statistical misadventures disclosed by the report. Further incentives are provided in sections 10 and 11 in which arguments that show that the failure of an investigator to deploy the appropriate statistical concepts, methods, and criteria is unethical as well as unscientific.

Although we have noted that most published studies fail to adequately assess the concordance either of model and data, or sample and prior estimates of parameters, etc., it must be granted that if it were subsequently found, on the evidence of statistically adequate analyses, that the model deployed in the study did actually fit the data and/or the a priori and sample information on the parameters (or on the response) were not mutually inconsistent, then the failure of the investigators to correctly address in their published report these crucial issues of concordance of the model estimates with the data and their consistency with prior beliefs would mean only that, although the documentation of the findings of the report was deficient, the reported findings would not actually misinform the reader.

However, our secondary analyses have disclosed that for each study for which the primary data were accessible to us, it is the case that, by statistically adequate measures, the received dose-response model fails to "fit" the data reported as empirical evidence for its validity either at all, or does not fit it as well - and by consequential amounts - as does the rival dose-response model of the same data. Either finding tends to diminish the confidence that we have in the interpretation assigned to the data by the original investigator. Our findings on the consistency of the estimates and inferences based on the received model with appropriate a priori information were similar. Thus, in these cases, the reported findings do misinform the reader.

Nisbett and Ross (1980) have remarked that it is sometimes the case that, "People's tendencies to persevere in their beliefs are so striking as to raise the possibility that such perseverance serves goals that may be more fundamental and important than holding correct views of particular issues."[5] In the present context "prudence" may be one such goal. "Convenience" may be another. For example, the selection of a linear, non-threshold, model of a low-dose radiation response instead of a rival curvilinear, quasi-threshold, model, on the criterion that the former is the more prudent signifies the presence of a motivational error in the belief system. For, as Raiffa (1982) has observed in his cogent discussion of risk assessment (assessment of uncertainties) and risk evaluation (policy analysis for risk management), "... there is a tension between honesty and prudence.: Probabilistic reports about adverse consequences to health are very often slanted to be conservative. I am arguing that it is better to report honestly, and that prudence should appropriately be represented in the evaluation process, not in the assessment process." (See also Herbert, 1986b).

There may be a similar conflict between validity and convenience. Fowler (1984) has argued that one of the reasons to believe in - or, at the very least, to deploy - the LQ model is its convenience: 1) "I shall use the $\alpha d + \beta d^2$ model because I believe it to be both valid for doses per fraction in the radiotherapy range, i.e., up to about 10 Gy, and convenient in application ..." 2) "Although the reciprocal total dose or F_e-plot (Douglas & Fowler, 1976) is not the most accurate way to calculate α/β it is the easiest method and gives a fairly accurate value if the data are good but only an optimistic estimate of the error range." (Fowler, 1989). 3) "The LQ model with the addition of the proliferation time factor, becomes more clumsy and dangerous than the simple LQ model used (correctly without a time factor) to calculate for late effects." (Fowler, 1989). 4) "The calculations outlined so far are extremely simple and can be done in one's head or on the back of an envelope - ..." (Fowler, 1989) (emphasis added). But none of these four arguments seem very cogent - even if the LQ model "fit" the data.

We must also note the comments by Toulmin (1970) on some of the sociological aspects of the formation and maintenance of a stable belief system in which he distinguishes between the roles of intellectual and magisterial authorities for belief and notes that, "... there is a tendency on the part of secondary workers in science to see only part of the intellectual picture in the subject with which they are concerned, and to restrict the choice of hypotheses by which they interpret their data, out of deference to the supposed example set them by a primary worker, whom they take as their master and whose magisterial authority they bow to." (We remark that "magisterial authority" may also be simply that of common practice, as defined by what gets published - see section 16.)

It is of no little interest here to observe that in the first of the six parts of his Opus Majus, published in the thirteenth century, the Franciscan friar, Roger Bacon, remarks on "... the four Universal Causes of All Human Ignorance - submission to faulty and unworthy authority, influence of custom, popular prejudice, and concealment of our own ignorance accompanied by an ostentatious display of our knowledge. Every man is entangled in these difficulties ..." (W. Whewell, 1856, has referred to R. Bacon's Opus Majus as the Novum Organum of the thirteenth century.)

As another instance of motivational error, one that we have already discussed, if a review of clinical trials includes examinations of only published studies then because of the prejudice against the null hypothesis, the conclusions and recommendations of the review may be biased toward "positive" findings. As remarked above, there have been several recent studies comparing the results of published and unpublished randomized control trials (RCTs) that have disclosed that RCTs which report "positive" findings are much more likely to be published than are RCTs reporting negative findings and recommend caution in the interpretation of published reports since the proportion of false positives is markedly inflated by the publication bias.

5.3.1 "Positive spin" in scientific papers?

The dis-incentives that lead to publication bias - "Publish or Perish" - are existential and insistent and it is therefore not at all surprising to find that there is considerable evidence to suggest that a variety of quite imaginative methodological innovations are not infrequently deployed by many investigators to impart a positive "spin" to the empirical evidence - data and/or statistics - of their studies. For instance, Williamson, et al (1986) report that the frequency of positive findings, say, rejection of the null hypothesis, in many scientific papers is inversely related to the adequacy of the methods used to obtain the reported results: "Approximately 80% of methodologically inadequate publications claimed positive findings, as compared with 25 per cent of methodologically adequate publications matched for research content." Evidently a possible motivation for study weaknesses may exist in the "career moves" of the investigators - as suggested in Kuhn's remark anent fruitfulness quoted above in section 3.2. Note also that in this aspect of modern scientific practice there is more than a little of the Aristotelian doctrine of final causes: the teleological explanation of an event or action in which, "... the goals or ends of an activity are dynamic agents in their own realization" (Kuhn, 1970a). This may help to explain the peculiarities of the design of some experiments and of the concomitant choice of a measure of concordance in some others - in those reports listed in Table 1.

The citations of the von Essen (1960) paper on volume effects in the subsequent authoritative reviews of this issue by Cohen (1982) and Withers (1988) provides additional evidence that the report of a positive finding in a scientific paper, such as that the association of two variables is significant, provides the empirical warrant of its "believability" even if it is obvious from a careful reading the paper that 1) the statistical test of the null hypothesis that is deployed therein is irrelevant to the matter at issue and 2) the statistic itself is incorrectly constructed from the data of the study. The von Essen paper is also a fair example of how an investigator may - however unwittingly - impart a "positive spin" to his results. Since this paper also provided the initial impetus to form the Task Group 1 and to embark on the series of secondary analyses that are described in Annexes I-IV, it merits a brief examination at this point. We shall also have occasion to briefly recur to it again, as providing an instance of "non-experiential data". The secondary analysis of this study is described in detail in Appendix I of this report.

In the von Essen study of volume effects, a chi-squared significance test, based on a 2x2 cross-classification of the data in a four-fold contingency table on a null hypothesis - no association - was used to interpret the study data on skin and tumor radiation responses. However, in the context of the ends to be achieved by the study, it is obviously the strength, not the significance, of the association of the two categories, that is crucial. Our secondary analysis of these data disclosed that although the sample size was large enough to reject the null hypothesis of no association, the strength of that association is trivial - and of no use whatever for the purposes that were to be served by the study. Moreover, it was evident from the examination of the procedures by which the two dichotomies of the 2x2 array were defined that the use of a test of hypothesis was not only irrelevant to the (scientific) purposes of the study, but was quite invalid for these data because the specification of one of the two categories was not given a priori, but instead was derived from the sample since it was systematically altered - as in a discriminant analysis procedure based on order statistics (Kendall, 1966) - in order to maximize the chi-squared statistic that summarized the table and thereby, of course, the significance of the association. The procedure followed by von Essen in the construction of his cross-classifications also led him to incorrectly identify the respective discriminant curves as the 0.03 (skin) or 0.99 (skin cancer) isoeffect curves. But a discriminant curve coincides with the 0.50 isoeffect curve (See Appendix I).

It appears to us that the only purpose that could be served by the reporting of what, in the event, proved to be spurious chi-squared hypothesis tests, was to strengthen the case for publishing the study by wrapping the results in a statistically significant, or "positive", finding - in a null hypothesis that could be "rejected" - although it required a bit of work.

It is the case, of course, that it is not always necessary to deploy innovations that impart a "positive spin" to the evidence. We have shown in section 5.1.2 that it is quite possible for the empirical evidence presented in the report of a study to be logically devastating to the "positive" conclusions of the report that offer that evidence in support of the findings and yet the report of the study is still published - and cited in authoritative reviews. This provides strong evidence for what may be referred to (borrowing from a recent inaugural lecture at the University College of London) as, "The Unimportance of Being Right."

6. "Should we believe it?"

"At times, ... the author ... of the research report bases the conclusions on 'evidence' that has come from a study characterized by an inadequate research design, improper procedures followed during the data collection phase of the study, or faulty statistical treatment of the data. One or more of these defects may mean that rival hypotheses exist that will lessen the confidence that can be placed in the researcher's explanation of what the data mean."

S. Huck and H. Sandler, 1979

"Nor is it, to be sure, necessary that these hypotheses be true, or even probable, but this one thing suffices, namely, whether the calculations show agreement with the observations".

A. Osiander, 1540

"The fundamental idea, and one that I should naturally accept, is that a law should not be accepted on data that themselves, show large departures from its predictions."

H. Jeffreys, 1961

"The pragmatic success of a scientific theory seems to guarantee the ultimate success of its associated explanatory mode."

T. Kuhn, 1977

"... the greater part of the variation is accounted for by the law, leaving an unexplained balance. The ground for accepting and generalizing the law is therefore a quantitative one: how much of the observed variation needs to be explained before we can accept the law?"

H. Jeffreys, 1960

"A relationship becomes law-like when different sets of data are summarized or modelled by the same quantitative equation. Its status depends upon the range of empirical conditions under which it holds".

A. Ehrenberg, 1975

"Theory is a policy - not a creed."

J. J. Thomson, 1912

The third question addressed by the Task Group devolves into a distinction between belief and knowledge. One difference between belief and knowledge is, of course, that knowledge is founded upon evidence; that is, knowledge is belief that can be rationally justified (Salmon, 1967). Assertions unsupported by evidence are merely dogma. Since knowledge is founded upon evidence it can be revised as the evidence changes. On the other hand, beliefs are much less susceptible to revision ("Magical thinking involves our inclination to seek and interpret connections between the events around us, together with our disinclination to revise belief after further observation" P. Diaconis, 1985). Thus, the answer to the third question posed by the task group was sought by a systematic statistical evaluation of the published evidence for each of the received models. Our anatomization of studies discloses the presence of some interesting pathology in the received modelling praxis of this field.

The evaluation of each of the studies listed in Table 1 that we reviewed proceeded from the following point of view: On the basis of a finite sample of observations the study has described, in terms of a parametric model (either empirical or mechanistic), a theoretical population to which the data are alleged to belong. The report must provide cogent answers to two fundamental questions (Baker and Nelder, 1978): a) Does the evidence of this sample support the assertion that the proposed model is a reasonable description of the population, i.e., that the model and sample refer to the same population? More precisely, does the set of n observations describe a random sample from the population described by the model? b) Since the parameters of the model are unknown, what are the "best values" and what are the respective ranges of values that the data suggest for them? We note here that, to be acceptable, the evidence for the answer to the first question must include both aggregate goodness-of-fit statistics and regression diagnostics and the evidence required for the answer to the second query must include both point and set (or interval) estimates of the parameters, statistical criteria that will be further described in Statistical methods II. section 7.

A special feature of these evaluations was the comparison, with respect to these questions, of the received model with a rival model of the same data. "If meaningful hypotheses ... entail specific statistical models ... then it should be possible to compare hypotheses by comparing the accuracy with which their implied models describe the data" (Walker and Rothman, 1982). Or as Kuhn (1970a) has remarked, "It makes a great deal of sense to ask which of two actual and competing theories fits the facts better." Therefore, for each study we offer a rival model of the study data, implying an alternative hypothesis, a different interpretation from that made by the original authors, for why their data turned out as they did.

It is, of course, the case, that in the comparison of two alternative models on a common set of data there are three possible alternatives: a) both models "fit"; b) neither model "fits"; c) one model "fits" and the other does not. In most comparisons that we made, case c) obtained: The rival model "fit" the better; frequently the received model did not "fit" at all - on the evidence of statistically adequate measures.

As remarked earlier, it is shown in Annexes I-IV that the evidence published for many of the hypotheses, estimates, and inferences that together comprise the received beliefs, resides in nothing more substantial than methodological artifact. Here we mention briefly a few of these; we will examine each in more detail - and in different contexts - later in this report. 1) The curvature, k, of a plane curve, $y = f(x)$, is defined as $k = (d^2y/dx^2)/[1 + (dy/dx)^2]^{3/2}$. For $k < 0$ the curve is concave-down; for $k > 0$, the curve is concave-up; $k = 0$, identifies a turning point, or shoulder. The negative curvature ($k < 0$) in the low-dose region ($0 \leq D \leq 2.0$ Gy) of the dose-response curve of the single fraction LQ model of radiation lethality is, for some data, obviously "induced" by the conventional semi-log plots on which cell-survival curves are usually displayed. On an arithmetic plot these curves are concave-up ($k > 0$) in this region - and thus do not describe a shoulder in the low dose region. 2) Some estimates of the α/β ratio obtained from the multifraction LQ model by the method of the so-called F_e-plot are dominated by the single-fraction observation at D/N = 15 Gy, a level of D/N well beyond the stipulated range (0 $\leq$ D/N $\leq$ 10 Gy) of validity of the multifraction LQ model. If this observation is deleted the estimates of α/β change by a factor of about 2. 3) The "straightness" of the F_e-plot for the multifraction LQ model, used as evidence of the validity of this model, is, for some data, a consequence of a design of the experiment: the straightness of the F_e-plot is guaranteed, whether the multifraction LQ model "fits" the data - on statistically adequate measures - or not. 4) The fit of the received linear non-threshold model of incidence of radiation-induced mammary neoplasia in female Sprague-Dawley rats exposed to Co^{60} gamma rays depends upon the fact that the observations at the lowest (D=0) and highest (D=500) doses were deleted from the data to which the model was fit (a practice stigmatized more than a century earlier by Charles Babbage (1830/1971) as "trimming"). If these two extreme observations are retained, the linear, non-threshold model is not valid - it does not "fit" on the evidence of statistically adequate measures and criteria.

As the mathematician and systems analyst J. Casti (1989) has remarked, "... in dealing with the idea of a natural system, we must necessarily touch upon some basic philosophical questions, of both an ontological and epistemological character." And the philosopher Kyburg (1970) has remarked that, "... statistics and epistemology are essentially one subject." Thus, although we have no desire to even appear profound, it will be useful to dichotomize the study weaknesses disclosed by our review into ontological weaknesses and epistemological weaknesses, after the rubric followed by the physician, E. Murphy, in his recent exposition of some of the philosophical issues entailed in the analysis and interpretation of experiments in biology and medicine (Murphy, 1982). The ontological weaknesses are demonstrated by the findings that either the kinds or levels of entities that emerge from the received theory and/or practices may not exist in any true scientific sense; for example, a dose-response effect that, on the basis of received theory, "should be" linear is found to be curvilinear when the data offered as evidence for linearity are examined more closely. An example of a weakness of this nature is disclosed by comparing Figs. 13a and 13b with Figs. 30 through 33. The epistemological weaknesses are disclosed by evidence of faulty arguments and false conclusions; e.g., a dose-response model that is reported to "fit" a given set of data on the basis of measures and criteria having the authority of common practice in the field (e.g., it is said that the LQ model "fits" if the corresponding F_e-plot is "straight". See Fowler, 1984) is found not to when assessed by statistically adequate methods and criteria (An example of such a weakness is described vividly below in Figs. 14-18).

6.1 Ontological weaknesses in received modelling praxis

"One of the most insidious and nefarious properties of scientific models is their tendency to take over, and sometimes supplant, reality."

E. Chargaff, 1963

"It seems entirely reasonable that the statistical model employed should reflect the nature of the scale of measurement."

McCullagh and Nelder, 1983

With the brief statistical background developed above in part 4 we can now proceed to identify and examine more closely what appear to us to be the principal inherent ontological weaknesses in the regression models of radiation dose-response that dominate the peer-reviewed literature. These are the failures of the received models to render an adequate account of the facts a) that the radiation response is inherently contingent, b) that radiation exposures, or treatments, are inherently multivariate, and c) that the a priori information on the model parameters that is frequently deployed by investigators in model construction is itself uncertain, or contingent, and, not infrequently, irrelevant, or inappropriate, as well. We shall elaborate on each of these weaknesses below.

a) The aleatory, or contingent, nature of the outcome to radiation treatment of a cancer patient is explicitly acknowledged in received practice in such oft repeated statements as that the fundamental goal of radiation therapy is to maximize the probability of achieving uncomplicated control of the disease. (For example, "The object of clinical radiation therapy is to obtain, for each patient, the maximum probability of cancer cure while minimizing the likelihood of significant normal tissue damage." R. Yaes, 1988) It is the case however, that neither the joint nor (even) the marginal probabilities of the occurrences of the high-dose events of clinical interest in normal and malignant tissues that are educed by the radiation treatment schedules now administered to achieve that expressed goal are accounted for by the currently received models such as the linear-quadratic, LQ. Indeed, such models do not formally, or explicitly, attempt to address - or even acknowledge - the problem of maximizing the probability of uncomplicated control.

Instead of constructing dose-response models of the marginal probabilities of occurrence of such high dose events as the ablation of tumor, or the necrosis of normal tissue, or of the yet more relevant joint probability of the concomitant occurrence of these effects as a function of dose - and other covariates - current efforts are still directed to the construction of isoeffect models of the dose required, for given levels of covariates such as the number of fractions N and elapsed time T, to elicit a fixed level of the marginal probability of occurrence, say π, of a specified event in either normal or neoplastic tissues; more often it is the former. It is, of course, quite apparent that isoeffect models of the occurrence of single events cannot be generalized to models of joint occurrence - although dose-response models of single events can be so generalized quite readily as we shall demonstrate below. (See also Appendix II). Thus, the current clinical intentions are not intersected by the received models of the intended (joint) effects. Moreover, the constant level of effect that is to be achieved in the clinic is rarely well-identified. It is usually the case that the effect is stipulated to be an extreme proportion, e.g., π = 0.05 ("tolerance of normal tissues") and π = 0.99 ("control of disease") but the confidence limits on the quantile of the dose, say D(π), required to educe it are never given. It is, of course, the case that these confidence limits on extreme quantiles are inherently quite large. Moreover, all of the experimental, animal, studies are directed toward obtaining conditional point estimates of the π = 0.50 quantile, or ED50, in an isoeffect study, a level of response that is quite remote from the levels of concern for either effect of clinical interest. But the ED50 is, of course, the level of response for which the dispersion of the estimate is least, i.e., the confidence limits are smallest, a fact that suggests that it may be reported by way of imparting a "positive spin" to the results, in these (rare) instances in which a measure of dispersion is included in the report. For example, compare Figs. 39a and 39b, which provide confidence limits on only ED50 and ESD50, with the cogent Figs. 42a and 42b, which provide confidence limits on the lower quantiles 0.05, 0.10, etc. Figures 39a and 39b were published, but the Figs. 42a and 42b are the more clinically relevant. (And since the latter disclose that the lower limb of the 0.95 confidence limit lies in the regions ED < 0 and ESD < 0, it is quite evident that both of these models are incorrect - it is impossible for any "dose" to be negative.)

This aversion of received modelling practice in high dose radiation therapy to deal with the inherent uncertainty in the occurrence of any of the sequelae to irradiation does not end there. The most characteristic feature of any radiobiological model are the uncertainties that encumber its development, validation, and deployment: uncertainty in the form of the model; uncertainty in the values of the parameters of the model; uncertainty in the values of functions of the parameters of the model - such as the predicted response. Although such uncertainties can be rigorously treated by correct statistical methods, such methods are, apparently, rarely employed in the production of the published reports of high-dose radiation effects. This indifference of the peer-group to a fundamental ontological feature of their field finds expression in the absence of such measures of uncertainty as the goodness-of-fit of a model (represented by the probability distribution of a goodness-of-fit statistic such as χ^2) and the precision of estimates of the model parameters, or functions of these parameters such as the response (represented by the (1-α) confidence limits on the parameters, or on the response).

Curiously enough, however, received practice in the modelling of low-dose radiation effects such as mutagenesis and carcinogenesis does not disclose the presence of similar weaknesses, or at least not to such a degree as in the received practice in the modelling of high dose effects. That is, the received models of low-dose effects 1) are invariably models of dose-response, and 2) probabilistic measures of goodness-of-fit of model and data and of uncertainty in the estimates of the model parameters - and functions of those parameters - appear to be found much more often in the reports of radiation mutagenesis and carcinogenesis studies than in reports of radiation toxicity and lethality studies. However, it should be remarked that in some reports in which interval estimates of the model parameters - or functions thereof - are presented, these estimates may be incorrect. For example, in Annex III, part 4, it is shown that the estimate of $Var(\alpha/\beta)$ presented in Table 5.1 of NCRP 64 (1980) is in fact incorrect: the term in $Cov(\alpha/\beta)$ was omitted from the calculation. (Anecdotally, this appears to be a not uncommon sort of error in such estimates.)

As Jeffreys (1961) has remarked, any statement of natural law must include what he has referred to as the "epistemological content"[6] of the law that is represented by the "unexplained variation" in the observed response. This is the random part, ϵ_i, which is, of course, as essential a part of the law as is the "predicted variation", or deterministic part, μ_i, and therefore any "... valid statement of the law must express it." The "unexplained variation" in the observed response is, of course, "explained" by finding the form of the distribution of the ϵ_i - or at least a parametric form and the first two central moments of that distribution. Many physicists - and others trained in the deterministic disciplines - seem to believe that all of the variation in the observed response is adequately accounted for, or "explained", by the set of predictor variables given by theory. (If there is no unexplained variation then there is no need to describe it - and so it isn't described. Therefore, interval estimates, goodness-of-fit tests, etc., are often not included in the published accounts of the dose-response studies by these investigators.)

The distributions of the random part of an observed response can be conveniently described as either Normal, that is, Gaussian, or non-Normal, usually either Binomial or Poisson. To a good first approximation the nature of the distribution of ϵ_i - and hence of the response, y_i - can be determined "by inspection" from the metric of the response variable: If the response is a measurement, the distribution can be assumed to be Normal; if a count, then it can be assumed to be Poisson; if a proportion, then Binomial, etc. Moreover, to a first approximation, it can be further assumed that the data are homogeneous. That is, that the observed response at a given level of $\underline{x}_i^T$ is drawn from a single (conditional) distribution, e.g., Normal, Binomial, or Poisson.

Of course, the subsequent statistical analysis of the data, in particular, a systematic comparison of the observed response with that estimated from the model at each of the n levels of the predictor variables, $\underline{x}_i^T$, $1 \leq i \leq n$, i.e., a residual analysis, may well disclose that the distribution of the response is, instead, logNormal rather than Normal; or Negative Binomial rather than Poisson; or Beta Binomial rather than Binomial; and, moreover, that the observed (conditional) variation in response is not due altogether to random sampling from the homogeneous population postulated by the model but rather that the responses in the study population are heterogeneous - "contaminated". There is, of course, the important logical distinction between failing to reject homogeneity in a hypothesis test and proving that it is present. Moreover, it is frequently difficult to decide whether the evidence for the putative heterogeneity is due to contamination of the sample, a possible error in data acquisition, or to the incorrect identification of the frequency distribution of the response in a homogeneous population, a possible error in model specification.
b) The inherently multivariate nature of the deterministic part, μ_i, of the observed response, y_i, is ignored by most investigators although it is usually obvious in their data; that is, the response is a function of the joint effects of correlated predictor variables and covariates: dose, fractions, time, anatomical site, stage of disease, etc. Multivariate regression models of dose-response data require matrix calculus for concise and fruitful representation and conceptualization as well as for efficient data analysis. Indeed, "[Matrix algebra] is becoming as necessary to science today as elementary calculus has been for generations ... quantitative scientists need to have as part of their mathematical repertoire, an understanding of matrix algebra" (S. Searle, 1982).

We must here present briefly some examples of the evidence for our conclusion that the inherently multivariate nature of the deterministic part of the response is regularly ignored by most investigators. It will be recalled that the first models of radiation dose-response to be widely deployed were the isoeffect models constructed by Cohen (1969) from his collation of three sets of clinical data - some obtained as much as 20 years earlier. These models could be written as $D(\pi) = 10^aT^b$ or $\log D(\pi) = a+b \log T$, where $D(\pi)$ is the total dose that elicits a specified response in $100\pi\%$ of those subjects irradiated over T days. (Note that the form of the equation is $y = a+bx$, a simple univariate model.) Estimates of a and b were obtained by Least Squares methods. Subsequently, Ellis augmented the model - by a thoroughly ad hoc maneuver - to include the effects of fractionation, N, to give the revised, NSD, isoeffect equation: $\log D(\pi) = a+b\log T + c \log N$. Note that the received estimates of a(=log1800) for connective tissues b(=0.11) and c(=0.24) were not determined by any statistically adequate method but instead were determined from the introspections of Dr. Ellis. Moreover, a numerical level, say π^* of the level of the effect, $0 \leq \pi \leq 1$, was never explicitly specified. Instead, "π = tolerance", was implicit in the clinical deployment of the model where "tolerance" is a judgment call of the physician and is thus physician-specific. However, it is widely believed that $\pi^* \cong 0.05$. See Appendix II for a discussion of the several meanings of "tolerance".

It is the case, of course, that in samples of clinical isoeffect data the explanatory variables N and T are highly correlated. This effect greatly weakens the sample information on the respective weights, b and c assigned to T and N separately in determining the corresponding level of dose, $D(\pi)$, that is required to educe a specified level of response, π. Stronger non-sample information must be combined with sample information in order to obtain more defensible estimates of b and c. Statistically adequate multivariate methods for the construction of such combinations are described below in section 7.2, e.g., Ridge regression. These so-called post-hoc salvage methods provide more defensible estimates of the model parameters b and c than the ad hoc methods of Ellis.

More recently, an alternative isoeffect model, the multifraction LQ, has been developed by Fowler from the cognate model of cell survival data, $S = \exp(\alpha d + \beta d^2)$. Fowler has strenuously argued its superiority to the NSD (and also to the latter's progeny, the TDF and CRE) as a model of clinical response. For most of its existence the LQ model has been described by the rather curious isoeffect (F_e-plot) equation, $D(\pi)^{-1}(\pi) = (\alpha/E) + (\beta/E)D(\pi)N^{-1}$, where $D(\pi)$ and N are as defined above, and $E = -\ln S$, where S is (most often) the unspecified and unobservable level of survival at dose $D(\pi)$ in an undefinable cell population. Note that the isoeffect model (F_e-plot) has the simple form $y = a+bx$, as was also the case for the earliest clinical isoeffect equation. (There are some obvious problems with this form but we defer comment until later.) Note also that estimates of (α/E) and (β/E) - but not α and β - can be obtained by Least Squares methods. Most recently (~1988) the model has been augmented - again by a thoroughly ad hoc maneuver - to include the effects of time, T. This gives the revised LQ model (Fowler, 1989). The generalization of the above isoeffect model has the form $D^{-1}(\pi) = (\alpha/E) + (\beta/E)D(\pi)N^{-1} - (\gamma/E)TD^{-1}(\pi)$, an isoeffect equation still more curious than the isoeffect equation of the LQ model from which T was omitted. The respective cognate dose-response (cell-survival) equations are

$$-\ln S = \alpha D + \beta(D^2/N)$$

and

$$-\ln S = \alpha D + \beta(D^2/N) - \gamma T.$$

There are, of course, still other difficulties, as remarked above, i.e., there is considerable uncertainty concerning both the identity of the cell population and the level of survival to which the response, lnS, refers. Since neither the level of the response, S, nor the system in which it occurs can be either specified, or even defined, the received model of the LQ hypothesis has no empirical content. Therefore, even if the parameters, α, β, γ, of the dose-response equations could be estimated, they cannot be defined. Obviously, there are intractable conceptual and computational difficulties with both the dose-response and isoeffect models of the LQ hypothesis. Moreover, although the LQ model was developed from sample experimental observations on the response of animals irradiated at each of a set of levels of D, N, and T, the design of the experiments were such that the correlation of N and T in the sample was as large as in the non-experimental clinical samples. Thus, the sample information on the separate effects of N and T on the level of response was no stronger in the experimental data than in the non-experimental data.

Thus, the first (circa 1960) models of radiation toxicity initially included only variables representing the effects of dose, D, and time, T, on the response, or, more precisely, the effects of T on the dose, D(π), required to elicit a poorly specified level of response, π, $0 \leq \pi \leq 1.0$ in the irradiated tissues. (Note that the specification of the level of response, here represented by π, has always been quite ambiguous.) This model was subsequently augmented, by sui generis concepts and methods, to include the effects of the number of fractions, N, on D(π).

The more recent (circa 1980) linear-quadratic, LQ, model of radiation toxicity which is now advocated by many as a replacement to the previous model has undergone a similar development. As initially set forth it included only variables representing the effects of dose D and fractions N on the response, or more precisely, the effects of N on the dose, D(S), that is required to elicit an unspecified level of survival, S, in an indefinable cell population in the irradiated tissues. In practice, it is assumed that an observed response π $(0 \leq \pi \leq 1)$ can be used as a proxy measure of S, and although α and β could not be estimated separately, the ratio $\theta = \alpha/\beta$, could be estimated from isoeffect data. Recently, the LQ model has been augmented by sui generis concepts and methods to include the effects of time T on the response - but still a response that usually cannot be either defined or observed.

In summary, the early models of radiation toxicity, a multivariate response in which the levels are determined by the joint effects of the concomitant levels of D, T, and N, initially included D and T and subsequently added N. The more recent models initially included D and N and have subsequently added T. In each model the subsequent addition was "ad hockery". In neither model was the response either unambiguously specified (neither in nature nor in level) or readily observable (although the earlier models were less ambiguous than the more recent, in this latter feature). In neither model was the inherent multivariate nature of the deterministic part of the response correctly appreciated or usefully represented. To do either requires matrix calculus - as will be shown below.

As Bacon (1620) has remarked, "The ill and unfit choice of words wonderfully obstructs the understanding". And so, also, does the "ill and unfit choice" of a mathematical representations for the deterministic and/or stochastic parts of the observed response - as we have discovered, repeatedly, in our assessments of the papers and reviews listed in Table 1.

c) Non-sample, or a priori, information on the parameter vector $\underline{\beta}$ of a model which is either quite ambiguous or irrelevant, or both, is frequently deployed by investigators in the construction of radiobiological models. However, like such deployments themselves, the precision and relevance of such information is not demonstrated - nor even acknowledged. We give two examples:

i) The received models of samples of size n of laboratory observations on mutagenesis and cell survival are constructed from data that have been "normalized" by either subtracting or dividing, respectively, the observed responses, say m_i at doses $D_i > 0$, $2 \leq i \leq n$, by that at $D_1 \equiv 0$, say m_1. But it can be shown (see Annex IV, part 6) that this is equivalent to constraining the coefficient, β_0, of the term D_0, i.e., the intercept, to be equal to a function of the single observed response, m_1 at $D_1 = 0$. For the models of cell-survival $\beta_0 \equiv \ln(m_1)$. However, this implies that the weight, w_1 of the observation at zero dose is infinite, say $w_1 = 10^6$ at $D_1 = 0$ (equivalently, the variance, ψ, of the a priori information in the cognate constraint $\underline{r} = R\underline{\beta} + \underline{v}$, $Var(\underline{v}) = \psi$, is zero; see section 7.2.4 below).

ii) In the BEIR III report (1980) the sample estimates of the parameter vector $\underline{\beta}$ for the models LQ-L, L-L, and Q-L of non-leukemia cancer mortality were quite unstable, e.g., $\bar{\beta}_j/Var(\hat{\beta}_j) < 1$. $0 \leq j \leq 4$ (Table V-9 of the BEIR III report). They were stabilized (Table V-11 of the BEIR III report) by imposing constraints representing a priori, or non-sample, information obtained as ratios (say $r_1 = \beta 1/\beta 2$) of the parameters of the cognate models of leukemia incidence (Table V-8 of the BEIR III report) under the assumption that the covariance matrix, ψ, of the constraint was the null matrix, say $\psi = 10^{-6}I_2$, where I_2 is the 2*2 identity matrix. (See section 7.2.4 below). But it is evident from Table V-8 itself that this assumption is untenable since the estimates of the parameters of the cognate leukemia models are themselves quite unstable. And the relevance of the information conveyed by a dose-response model of leukemia (a non-epithelial tumor) incidence to a dose-response model of non-leukemia cancer (epithelial tumors) mortality is, of course, questionable (See section 7.11.2 and Annex IV, part 5).

Note that example i) above describes an instance in which non-sample information on the parameter vector, $\underline{\beta}$, of the model is inadvertently "mixed" with the sample information thereon by the uninformed deployment of improper, sui generis, data-analytic procedures that are equivalent to the Mixed estimation procedures to be discussed in section 7.2.4. (The "mixing" in such instances might better be called "contamination".) Example ii) describes an instance of the deliberate deployment of such non-sample information in order to "stabilize" the sample estimates of the parameters of the model, that is, to reduce the overall uncertainty in the parameter estimates (It is demonstrated in section 7.2.4 below, that the variance of the parameter estimate is reduced from $Var(\hat{\underline{\beta}})$ to $Var(\hat{\underline{\beta}}^{**}) = [Var(\hat{\underline{\beta}})^{-1} + R^T\psi^{-1}R]^{-1}$ by including non-sample information on $\underline{\beta}$ in the form of the constraint $\underline{r} = R\underline{\beta} + \underline{v}$, $E(\underline{v}) = \underline{0}$, $Var(\underline{v}) = \psi$.)

It is evident that in either example, the failure of the investigator to correctly describe the level of uncertainty in the non-sample, or a priori, information that is represented by ψ, will lead to biased estimates of the accuracy and precision of the information on the parameter vector, $\underline{\beta}$, of the model.

6.1.1 Multivariate treatments and responses

"At a practical level, what distinguishes techniques considered univariate from those considered multivariate is whether matrix manipulations are required."

R. Harris, 1975

Multivariate data are represented by an observation matrix, [y, X], a two-dimensional (n*k) array in which the n rows are the observations (e.g., patients, animals, etc.) and the k columns are variates (e.g., response, dose, fractions, age, sex, etc.), each an (n*1) vector. One of the variates, y, is the response. The remaining k-1 = p columns are the covariates of the X matrix. The response variable may be either discrete (counts, fractions) or continuous (measurements). The covariates may be either qualitative or quantitative. The quantitative covariates are referred to as measurements, e.g., dose, time, etc. The qualitative covariates are factors, e.g., stage, sex, etc. The observation matrix may be written explicitly as,

$$[\underline{y}, X] = \begin{bmatrix} y_1, & x_{11}, & x_{12}, & \ldots, & x_{1p} \\ y_2, & x_{21}, & x_{22}, & \ldots, & x_{2p} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ y_n, & x_{n1}, & x_{n2}, & \ldots, & x_{np} \end{bmatrix}$$

$\underline{y}$ is the response vector, e.g., y_1 may be the # responders in the "first" observation (row), etc. X is the design matrix, e.g., $x_{11}, x_{12}, \ldots, x_{1p}$ may be the dose level, # fractions, ... etc., in the first observation (row), etc. Thus, the i^{th} row is the observation $(y_i, \underline{x}_i^T)$. It has been found that regression models of dose-response are most easily and concisely developed, assessed, and deployed by the use of matrix methods. Indeed, all computer algorithms for regression analysis are simply a concatenation of elementary matrix operations and manipulations.

Matrices are manipulated according to the concepts and rules of matrix algebra. Two brief and elementary textbooks on matrices are Matrix Theory for Physicists by J. Heading (Longmans, Green. NY. 1958) and A Modern Algebra for Biologists by H. Nahikian (Univ. Chicago, Chicago. 1964). Both are classics. Another classic is Applied Group-Theoretic and Matrix Methods by B. Higman (Clarenden Press, Oxford, 1955). A good recent textbook is Matrix Algebra Useful for Statistics by S. Searle (J. Wiley. NY. 1982).

We present below by way of review a brief exposition of elementary matrix operations and manipulations based upon these four texts as well as a set of useful alternative equivalent notations for linear regression models.

6.1.1a Review of matrix notations and operations

A matrix is a two-dimensional array of numbers. An (mxn) matrix, say X, consists in m rows and n columns of numbers:

$$X = \begin{pmatrix} x_{11}, & ..., & x_{1n} \\ ..., & x_{kl}, & ... \\ x_{ml}, & ..., & x_{mn} \end{pmatrix}; \text{ say, } X = \begin{pmatrix} 26, & ..., & 7 \\ ..., & 15, & ... \\ 11, & ..., & 132 \end{pmatrix}$$

The element, x_{ij}, of X is that number in the i^{th} row and j^{th} column of the matrix X where $1 \le i \le m$; $1 \le j \le n$. (The x_{ij} may be > 0, = 0, or < 0.)

A (1x1) matrix, say s, is a single number (often called a scalar). An (mx1) matrix, say x, is a single column of m rows (often called a column vector). A (1xm) matrix, say x^T, is a single row of m columns (often called a row vector). x^T is the transpose of x (vide infra).

The sum of two (1xm) matrices, say x^T and y^T, with elements x_{ij} and y_{ij}, respectively, is the (1*m) matrix z with elements $z_{ij} = x_{ij} + y_{ij}$. For example, $x^T = (1, 4, 7) = y^T = (3, 16, 7)$. Then $z^T = x^T + y^T = (4, 20, 14)$.

The inner product, say x^T*x of a (1xm) matrix x^T and an (mx1) matrix, x, is the (1x1) matrix, a scalar, say, s^2. $x^T*x = (1, 4, 7)\begin{pmatrix}1\\4\\7\end{pmatrix} = 1^2 + 4^2 + 7^2 = 66 = \Sigma x_j^2 = s^2$.

The operation may be generalized to (mxn), etc., matrices.

The outer (or dyadic) product, say $x*x^T$, of an (mx1) matrix x and a (1xm) matrix x^T is the mxm (square) matrix, say, X.

$$x*x^T = \begin{pmatrix}1\\4\\7\end{pmatrix}(1, 4, 7) = X = \begin{pmatrix} 1, & 4, & 7 \\ 4, & 16, & 28 \\ 7, & 28, & 49 \end{pmatrix}$$

The operation may be generalized to (mxn), etc., matrices.

The sum of two (m*n) matrices, say X and Y, with elements x_{ij} and y_{ij}, respectively, is the (m*n) matrix, say Z, with elements $z_{ij} = x_{ij} + y_{ij}$. For example,

$$X = \begin{pmatrix} 1, & 7 \\ 4, & 4 \\ 7, & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} 16, & 4 \\ 49, & 7 \\ 28, & 1 \end{pmatrix}, \quad Z = \begin{pmatrix} 17, & 11 \\ 53, & 11 \\ 35, & 2 \end{pmatrix}$$

The product of an (mxh) matrix, say X, and an (hxn) matrix, say Y, is the (mxn) matrix, say Z, where

$$z_{ik} = \Sigma y_{ij}x_{jk} \quad (1 < i < m;\ 1 < k < n)$$

For example,

$$X = \begin{pmatrix} 1, & 7 \\ 4, & 4 \\ 7, & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} 16, & 49, & 28, & 4 \\ 4, & 7, & 1, & 16 \end{pmatrix}, \quad Z = \begin{pmatrix} 44, & 98, & 35, & 116 \\ 80, & 224, & 116, & 80 \\ 116, & 350, & 197, & 44 \end{pmatrix} = X*Y$$

The inverse of an (mxm) symmetric matrix, say X, is a matrix, X^{-1}, such that $X*X^{-1} = X^{-1}*X = I_m$ where I_m is the (mxm) identity matrix, I_m = Diag[1]. For example,

$$X = \begin{pmatrix} 1, & 4 \\ 4, & 7 \end{pmatrix}, \quad X^{-1} = \begin{pmatrix} -0.778, & 0.444 \\ 0.444, & -0.111 \end{pmatrix}, \qquad X*X^{-1} = I_2 = \begin{pmatrix} 1, & 0 \\ 0, & 1 \end{pmatrix}$$

The transpose of an (mxn) matrix Y is the (nxm) matrix Y^T (and Y is the transpose of Y^T).

The product of an (nxm) matrix and its (mxn) transpose is a symmetric (mxm) matrix, say, Y^T*Y. For example,

$$Y = \begin{pmatrix} 3, & 1, & 4 \\ 3, & 4, & 16 \\ 3, & 7, & 28 \\ 3, & 10, & 35 \end{pmatrix}, \quad Y^T = \begin{pmatrix} 3, & 3, & 3, & 3 \\ 1, & 4, & 7, & 10 \\ 4, & 16, & 28, & 35 \end{pmatrix} \text{ and } Y^TY = \begin{pmatrix} 36, & 66, & 249 \\ 66, & 166, & 614 \\ 249, & 614, & 2281 \end{pmatrix}$$

A square symmetric matrix can be represented as a linear combination of <u>outer products</u> (dyads):

$$X = \lambda_1\underline{v}_1\underline{v}_1^T + \lambda_2\underline{v}_2\underline{v}_2^T + \lambda_3\underline{v}_3\underline{v}_3^T = \Sigma\lambda_j\underline{v}_j\underline{v}_j^T$$

where λ_j is the j^{th} <u>eigenvalue</u> of the matrix X and v_j is the j^{th} <u>eigenvector</u> of that matrix. ($\Sigma\lambda_j\underline{v}_j*\underline{v}_j^T$ is called the <u>spectral decomposition</u> of X). Note that $\underline{v}_j^T\underline{v}_j = 1.0$. $1 \leq j \leq 3$.

The <u>Trace</u> of X is the <u>Sum</u> of the eigenvalues, λ_j:

Trace(X) = $\Sigma\lambda_j$. $1 \leq j \leq m$.

The <u>Determinant</u> of X is the <u>product</u> of the eigenvalues, λ_j:

Determinant (X) = $\Pi\lambda_j$. $1 \leq j \leq m$.

The <u>Condition Number</u>, κ, of an (mxm) matrix is a measure of the "spread" in the spectrum of eigenvalues of the matrix; it is the ratio of the largest, $\lambda_{(m)}$, to the <u>smallest</u>, $\lambda_{(1)}$, of the eigenvalues: $\kappa \equiv \lambda_{(m)}/\lambda_{(1)}$. The condition number of a symmetric matrix, Y^TY, is a measure of a "condition" of the system that is described by the matrix Y. Specifically, it is a measure of the degree of <u>multicollinearity</u> (linear dependence) present in the variables that are represented by the <u>columns</u> of Y. For example column 3 of Y (above) is nearly a multiple (4*) of column 2. The correlation coefficient of these two columns is 0.993. The <u>eigenvalues</u>, λ_j, and <u>eigenvectors</u>, $\underline{v}_j$, of the matrix Y^TY are,

$\lambda_1 = 8.9001$	$\underline{v}_1^T = (0.9871, -0.1441, -0.0692)$
$\lambda_2 = 0.5519$	$\underline{v}_2^T = (0.1205, 0.9552, -0.2703)$
$\lambda_3 = 2473.5480$	$\underline{v}_3^T = (0.1051, 0.2585, 0.9603)$

Then,

Trace (Y^TY) = <u>2483.00</u>

Determinant (Y^TY) = <u>12150.00</u>

Condition Number (Y^TY) = 4481.89

And, "Generally, if the condition number is less than 100, there is no serious problem with multicollinearity. Condition numbers between 100 and 1000 imply moderate to strong collinearity, and if κ exceeds 1000, severe multicollinearity is indicated" (Montgomery and Peck, 1982). We shall discuss below the effects of <u>multicollinearity</u> in X on the sample estimates of $\underline{\beta}$ for the regression model, $\underline{y} = X\underline{\beta} + \epsilon$.

6.1.1b. <u>Review of notation for a linear regression model</u>

A <u>simple</u> linear regression model may be written as

1) $y = a + bx + \epsilon$
2) $y = a_0 + a_1x + \epsilon$
3) $y_i = a_0 + a_1x_{1i} + \epsilon_i^{\#} = \Sigma a_k x_{ki} + \epsilon_i$, $(1 < i < n)$
where $x_{0i} = 1$
4) $y_i = \underline{x}_i^T\underline{a} + e_i$ $(1 \leq i \leq n)$
where $\underline{x}_i^T = (1, x)$, $\underline{a}^T = (a_0, a_1)$
5) $\underline{y} = X\underline{a} + \underline{\epsilon}$
where $\underline{y}$ is (nx1), X is (nxk), $\underline{a}$ is (kx1), $\underline{\epsilon}$ is (nx1), and here k = 2.

A <u>multiple</u> linear regression model may be written as

1) $y = a + bx + cz + \epsilon$
2) $y = a_0 + a_1x_1 + a_2x_2 + \epsilon$
3) $y_i = a_0 + a_1x_{1i} + a_2x_{2i} + \epsilon_i^{\#} = \Sigma a_k x_{ki}$, where $x_{0i} = 1$, $(1 \leq i \leq n)$
4) $y_i = \underline{x}_i^T\underline{a} + \epsilon_i.^{\#}$, $(1 \leq i \leq n)$
where $\underline{x}_i^T = (1, x_1, x_2)$, $\underline{a}^T = (a_0, a_1, a_2)$
5) $\underline{y} = X\underline{a} + \underline{\epsilon}$
where $\underline{y}$ is (nx1), X is (nxk), $\underline{a}$ is (kx1) and $\underline{\epsilon}$ is (nx1). k = 3.

[#] $(y_i, \underline{x}_i^T)$ is the i^{th} row in an (nxk) matrix.

In the present report, it will be found that there are two senses in which the response of an organism to a treatment must be described as inherently multivariate. In the first, more common, but perhaps less accurate, sense in which it will be encountered in this report, there is only one system, say Q_1, in the treated organism that exhibits a unique quantal, or binary, radiation response, say $[\bar{E}, E]$, that is of interest to the investigator, at each level of treatment, $\underline{x}^T$. In this sense, for a sample of size n, the observed response is described by the (n*1) column vector $\underline{y}$, with the n components $y_i = \mu_i + \epsilon_i$, $1 \leq i \leq n$. The deterministic part, μ_i, of each component, y_i, is a function - either linear or non-linear - of the (1*p) treatment vector, $\underline{x}_i^T$, where $p > 1$, and a vector of unknown parameters, $\underline{\beta}$, say $\mu_i = f(\underline{x}_i^T, \underline{\beta})$, where the components of $\underline{x}_i^T$ represent the radiation dose and additional covariates. In this sense, the response is termed multivariate because it includes a deterministic part, μ_i, which is a function of a multivariate treatment vector, $\underline{x}_i^T$, just as it is stochastic because it includes a stochastic part, ϵ_i. In this sense, for a generalized linear model of a multivariate quantal response, we have, $g(\mu_i) = z_i$, the probit transform, and $z_i = \eta_i = \underline{x}_i^T\underline{\beta}$, the linear predictor, where $z_i = \Phi^{-1}(\pi_i)$, $0 \leq \pi_i \leq 1$, and $\Phi(.)$ is the Normal distribution function. For a radiation treatment vector in which p=3, representing a total dose D that is delivered in N fractions over a span of time T, the linear predictor of a multivariate probit model is $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$, $1 \leq i \leq n$, where $x_1 = \log D$, $x_2 = \log N$, $x_3 = \log T$. A statistically adequate description of this multivariate response includes the sample estimates of the parameter vector $\underline{\beta}$, which describes the conditional (on N and T) distribution of the tolerance dose for the system Q. The probability of occurrence of the quantal response E in system S at $\underline{x}_i^T$ is just $\eta = \Phi(z_i)$. The set of treatment variables X_1, X_2, ..., in which X_j, $1 \leq j \leq k$, is an n*1 column vector, that comprise the matrix X, are, in general, correlated, as remarked above. One important consequence of this feature of the multivariate nature of the treatment to be remarked at once is that the value obtained for the sample estimate, $\hat{\beta}_j$, of the respective parameter, β_j, is conditional upon the presence or absence of each of the other treatment variables. This important effect does not appear to be widely appreciated in the literature. We examine it in more detail below.

In the second, more accurate, usage of the term, multivariate response, there is more than one system, say Q_1 and Q_2, in the treated organism, that exhibits a unique quantal response, of interest to the investigator, at each level of treatment, $\underline{x}^T$. For example, if the treatment is a drug, it usually affects more than one physiological system in the treated organism, producing say, a "main effect" in system Q_1 and a "side effect" in system Q_2. In this case, for a sample of size n, the observed response is described by the (n*2) matrix $Y = [\underline{y}_1, \underline{y}_2]$, with components, y_{ij}; $j = 1,2$; $1 \leq i \leq n$. In this sense, the response is multivariate because there is more than one response variate, each with a deterministic and a random component, although the treatment, $\underline{x}_i^T$, is not (necessarily) multivariate, i.e., it may be that p=1. For example, for a multivariate probit model, in this second sense, we may have the linear predictor, $\eta_{ij} = \beta_{0j} + \beta_{1j}x_i$; $j = 1,2$, $1 \leq i \leq n$.

A statistically adequate description includes sample estimates of the respective parameter vectors, $\underline{\beta}_1$ and $\underline{\beta}_2$, of the two systems, Q_1 and Q_2 together with the estimate of the correlation coefficient, ρ_{12}, measuring the association between the respective tolerances of systems Q_1 and Q_2.

The probability of occurrence of the quantal response, say E_j, for system Q_j at $\underline{x}_i^T$ is

$$\pi_j^1(\underline{x}_i^T) = \Phi_1(z_i)$$

and the probability of occurrence of the complementary response, $\bar{E}_j$, is

$$\pi_j^0(\underline{x}_i^T) = 1 - \Phi_1(z_i)$$

where $z_i = \underline{x}_i^T\underline{\hat{\beta}}$. Note that these are conditional probabilities; that is, the probability is a function of the treatment vector, $\underline{x}_i^T$. When the joint response of two systems is of interest, it is necessary to introduce the joint distribution of the respective tolerances. To generalize the univariate probit model, it is necessary to take this joint distribution to be the standardized multivariate Normal distribution, $\Phi_2(z^1, z^2; \rho_{12})$. Then the probability that both Q_1 and Q_2 exhibit their respective quantal responses at $\underline{x}_i^T$ - the event $[E_1, E_2]$ - is

$$\pi_{12}^{11}(\underline{x}_i^T) = \Phi_2(z_i^1, z_i^2; \rho_{12}).$$

The probability that Q_1 manifests the response and Q_2 does not - the event $[E_1, \bar{E}_2]$ - is

$$\pi_{12}^{10}(\underline{x}_i^T) = \pi_1^1(\underline{x}_i^T) - \pi_{12}^{11}(\underline{x}_i^T).$$

Similarly, the joint probability of the event $[\bar{E}_1, E_2]$ is

$$\pi_{12}^{01}(\underline{x}_i^T) = \pi_2^1(\underline{x}_i^T) - \pi_{12}^{11}(\underline{x}_i^T).$$

The probability that neither system responses at $\underline{x}_i^T$, the event $[\bar{E}_1, \bar{E}_2]$, is

$$\pi_{12}^{00}(\underline{x}_i^T) = 1-\pi_{12}^{10}(\underline{x}_i^T) - \pi_{12}^{01}(\underline{x}_i^T - \pi_{12}^{11}(\underline{x}_i^T).$$

Similar equations hold for the joint response more than two systems.

A treatment success, the event, S, in radiation therapy can be described as the occurrence of the compound event, uncomplicated control of disease. That is, it is the joint occurrence of the events, E_1, ablation of tumor, and $\bar{E}_2$, absence of the treatment-limiting event, E_2, in the normal tissues of the target volume, or $S = E_1$ and $\bar{E}_2$. Similarly, a treatment failure is the event $\bar{S} = \bar{E}_1$ or E_2. Obviously, $P(\bar{S}) + P(S) = 1$. In the above notation we have $P(E_1$ and $\bar{E}_2) = \pi_{12}^{10}$ and $P(\bar{E}_1$ or $E_2) = \pi_{12}^{11} + \pi_{12}^{01} + \pi_{12}^{00}$. Note that this formulation of these events provides a model for the prescription of those levels of dose, fractions, and time that will achieve a specified probability of uncomplicated control, P(S). In an obvious notation we have to find the solution to the set of three simultaneous equations:

$$z_1 = \beta_{10} + \beta_{11}\log D + \beta_{12}\log N + \beta_{13}\log T$$
$$z_2 = \beta_{20} + \beta_{21}\log D + \beta_{22}\log N + \beta_{23}\log T$$
$$\log(5/7) = \log N - \log T$$

where z_1 and z_2 are the probit transforms of $P(E_1)$ and $P(E_2)$ respectively, and the third equation represents the constraint, $N = (5/7)T$, for the usual treatment regimen of 5X per week. The sample estimates of the respective parameter vectors, $\underline{\beta}_1$ and $\underline{\beta}_2$, are obtained by the multivariate probit methods of Lasaffre and Molenberghs (1991). See also Herbert (1993b).

As described in section 6.1a) above, in clinical radiation biology the investigator is interested in models of organisms that include three systems, say Q_1, Q_2 and Q_3 for which the observation matrix is [Y, X] where both the response, Y, and the treatment, X, are multivariate. That is, the response is represented by the matrix $Y = [\underline{y}_1, \underline{y}_2, \underline{y}_3]$ in which $\underline{y}_1$, $\underline{y}_2$ and $\underline{y}_3$ represent the unique quantal response vectors for ablation of tumor, "early" reactions of normal tissues, and "late" reactions of normal tissues, respectively. The treatment is represented by the matrix $X = [X_1, X_2, X_3]$ where X_1, X_2 and X_3 are, respectively, functions - which may include the identity but is often the logarithm - of the treatment variables, dose, fractions and time. In this representation it is both useful and possible to describe the relationships between the responses of the different systems of the irradiated organism in terms of the correlation structure of the matrix Y as well as the dose-response relation, $g(\mu_j) = \eta_j$, $1 \leq j \leq 3$, for each separate system; i.e., in terms of the joint and marginal probabilities described above.

Of course, the correlation structure of the matrix X must also be correctly represented in any account of either the joint or marginal responses.

It should be noted that, to date, dose-response data - either experimental or non-experimental - in which the response of the sample can be completely specified by a (n*1) vector, $\underline{y}$, that is a function of the matrix, $\underline{X}$, are far more common in the literature than those in which the response of interest must be described by a (n*3) matrix, Y, although the latter response would appear to be of no little interest to radiation oncology. There is, however, very little data in which the concomitant occurrence of unique quantal responses in even two systems of an irradiated organism, e.g., ablation of tumor and necrosis of the normal tissues in which it is embedded, lying within the same target volume, is recorded. Among the studies listed in Table 1, only von Essen (1960) described clinical data on the joint response of two systems. von Essen (1960) presents a super-position of two families of isoeffect curves each indexed by the size of the volume of the irradiated tissues. One family describes tumor ablation; the other describes the concomitant occurrence of necrosis in normal tissues within the same target volume.

In Appendix II we describe the construction of a model of the probability of the joint occurrence of necrosis of normal tissues and ablation of tumor for a sample of clinical observations on the radiation treatment of patients with cancer of the oropharynx. See also Herbert, 1993b.

6.1.2 Under-fitting and over-fitting models

"... both improper omission and indiscriminate inclusion of variables lead to compromised inferences."

J. Robins and S. Greenland, 1986

Obviously, a dose-response model can be misspecified either by the incorrect choice of the functional form of the deterministic part of the response, μ_i, or by the incorrect choice of the form of the distribution of the stochastic part of the response, ϵ_i, (or both). In most of those studies listed in Table 1 the models that were discussed, deployed, etc., were found to be misspecified in both the deterministic and stochastic parts of the response. DuMouchel and Harris (1983) have remarked, that, "... there is a formal connection between errors of model misspecification within individual experiments and errors of extrapolation between experiments." Since extrapolated estimates of response are often required of the dose-response models in Table 1, their remark suggests that investigators should be a bit more diffident toward extrapolation of these models than has hitherto often been the case.

In Annexes II-IV we demonstrate the ways in which the received models and data fail to intersect in the most important features of any radiation response: the specifications of the deterministic and stochastic parts of the response - and the profound effects of these misspecifications on both the bias and precision of the estimates and inferences made from these models - even in interpolation. Two common ways in which the deterministic part of the generalized linear model may be misspecified are termed overfitting and underfitting.

In overfitting, more predictor variables are included in the linear predictor, η_i, than are required to obtain a statistically adequate fit to the observed data with the result that the variances of the parameter estimates and of linear functions of the parameter estimates such as the response, $\underline{x}_i^T\hat{\underline{\beta}}$, and of non-linear functions of the parameter estimates such as $\bar{\theta} = \hat{\beta}_j/\hat{\beta}_k$ are inflated by the presence of the redundant variables.

An egregious example of overfitting is provided by the BEIR III models of the LSS(T65D) data that are described in Tables V-8 and V-9 of BEIR III (1980). Each of the six models includes a (0,1) indicator variable for city (Hiroshima and Nagasaki) for which the sample estimate of the coefficient is much less than its standard error. The presence of this non-significant "city-variable" further weakens the already weak (imprecise) parameter estimates. (See section 7.1 and also Herbert, 1986c).

A vivid example of overfitting leukemia incidence data by the LQ model is shown in Fig. 21a. Moreover, in the case of the latter, the bias, $E(\hat{\theta}) - \theta$, as well as the variance, of $\theta = \beta_1/\beta_2 = \alpha/\beta$, is inflated by overfitting. The inflation of the bias of the estimate, $\bar{\theta}$, as $Var(\hat{\underline{\beta}})$ is inflated, does not seem to be widely appreciated although it can be quite large as is evident from the so-called delta approximation:

$$E(\bar{\theta}) - \theta = \beta_1 Var(\hat{\beta}_2)/\beta_2^3 - Cov(\hat{\beta}_1, \bar{\beta}_2)/\beta_2^2$$

See section 7.3 and Annex III, part 4 for a discussion and examples. This is one pragmatic basis for the desideratum of parsimony in the construction of hypotheses and models that is expressed in the principles of Ockhams' Razor[7] and Mach's Economy of Thought[8]. Another argument for parsimony was given by Jeffreys: ... the simplest law is chosen because it is most likely to give correct predictions ..." Thus, the more parsimonious model should give estimates for which both the bias and variance are minimal. However, as Robins and Greenland (1986) remark, "... any parsimony principle is irrelevant ..." for a predictor variable for which there is strong a priori evidence for including it in the model. Indeed, there can be an excess of parsimony: the model can be underfit.

In underfitting, important predictor variables are omitted leading to biased (aliased) estimates of the parameter $\underline{\beta}$ and of functions of the parameter $f(\underline{\beta})$. See Annex II, part 3 and 5 for discussion and examples. A vivid example of underfitting cell-survival data by the LQ model is presented in Fig. 20b. Since ample discussion of underfitting, as well as demonstrations of underfitting, are provided in Annex II, parts 3 and 5, we give only a brief description of the effect at this point. Assume that the correct model of a given set of observations is linear with parameter vector $\underline{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3)$ and that the observed response has a Normal distribution. The model can be written as $\underline{y} = X\underline{\beta} + \underline{\epsilon}$. Assume now that these observations are underfit by the linear model $\underline{y} = X\underline{\beta}^* + \underline{\epsilon}^*$ where $\underline{\beta}^{*T} = (\beta_0^*, \beta_1^*, \beta_2^*)$. Then the least-squares estimate of $\underline{\beta}^*$ is biased (aliased) by an amount that depends on the relation of the omitted variable, say X_3, to the response, $\underline{y}$, and also on the distribution of the observations on X in the original sample:

$$\begin{pmatrix}\beta_{0*}^{*}\\ \beta_{1*}\\ \beta_2\end{pmatrix} = \begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} + \beta_3\begin{pmatrix}A_0\\ A_1\\ A_2\end{pmatrix}$$

where $A = (X^TX)^{-1}X^TX_3$ is the so-called alias matrix. It is apparent from the form of the matrix A that, "The bias terms ... depend not only on the postulated and true models but also on the experimental design ..." (Draper and Smith, 1981). Since the alias matrix is always study-specific, the reported parameter estimates $\hat{\underline{\beta}}^*$ - and functions thereof such as $f(\underline{\beta}^*) = \alpha/\beta$ - may be highly idiosyncratic, thereby making comparison of estimates of $\underline{\beta}$ and/or $f(\underline{\beta})$ between studies very difficult. Moreover, owing to the error of misspecification described above - underfitting - the errors of extrapolation between experiments can be quite large - as remarked by DuMouchel and Harris (1983). It is apparent that, "... both improper omission and indiscriminate inclusion of variables in a model can lead to compromised inferences", as remarked above.

6.1.3 Multicollinearity and non-uniformity in [y, X]

"... there is a special problem caused by collinearity. This is the problem of interpreting multidimensional evidence. Briefly, collinear data provide relatively good information about linear combinations of coefficients. The interpretation problem is the problem of deciding how to allocate that information to individual coefficients. This depends on prior information."

E. Leamer, 1978

As remarked above, one effect of ignoring the inherent multivariate nature of the response is that the estimates and inferences obtained from models constructed from either experimental or non-experimental data, by any analytical methods whatever, are often biased and imprecise. A second consequential result of ignoring the multivariate nature of the response - in either of the two senses defined in section 6.1.1 - is that the designs of many of the published dose-response experiments are often flawed in consequential ways.

With respect to the second effect, flawed experimental designs, it is important to remark that, as evidence of this effect, in the studies reviewed in Annexes II-IV, we found that the levels of non-uniformity - especially the presence of "gaps" in the marginal distribution of dose, or functions thereof - and of multicollinearity that are present in the distributions of observations, [y, X], in designed experimental (laboratory) studies are at least as high as in those that are found in non-experimental (clinical and epidemiological) studies. In the latter studies the presence of such weaknesses - non-uniformity and multicollinearity - in the data are nearly inevitable; in the former their presence is quite gratuitous. See Figs. 1 - 4, below.

The presence of non-uniformity in the distribution of observations, [y, X], signifies that a large fraction of the sample of observations are included in a small fraction of the range of dose in the sample, e.g., in one well-known experiment on radiation mutagenesis, 85% of the observations are included in the lowest 25% of the range of dose. See Annex III, part 5. In another experiment on radiation toxicity, 75% of the observations are included in the lowest 50% of the range of dose. See Annex III, part 3, 4, and 5. The distributions of the observations in these two experiments are described in Figs. 1 and 2, respectively. The degree of non-uniformity is quite striking in each case. Still another instance of non-uniformity can be seen in the set of pooled dose-response experiments described in Tucker and Thames (1983). This set included seven dose-response curves in which the number of levels of dose for each of the seven levels of N and T varied between two and four. Since there were different numbers of observations at each level

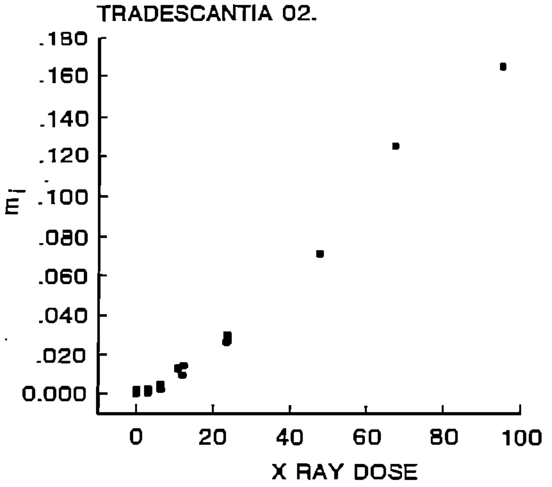


Fig. 1a. Scattergram of the observations, (m_i, D_i), $1 \leq i \leq n = 20$, for Sparrow et al (1972) observations on incidence of pink mutants in Tradescantia 02 over the range $0 \leq D_i \leq 100$ rad. m_i is the Poisson rate constant in mutations/stamen hair $* 10^5$ and D_i is the X-radiation dose in rads. Note i) the non-uniform distribution of D_i and ii) the non-uniform replication of observations at each level of D_i: The number of observations at each level of dose varies between 1 and 4; 85% of the observations are included in the lowest 25% of the range of dose (0-100 rad) over which the LQ model of mutagenesis is stipulated to be valid. In any model of these data the estimates and inferences may be driven by either those observations that are most replicated or those observations which are most remote. It is instructive to view this study as a meta-analysis in which the data of four weak experiments (#2, #5, #6, and #7 in the original report of Sparrow et al, 1972) were pooled. In only one of the experiments (#5) did the range of dose cover the stipulated range of validity (0-100 rad) of the LQ model. The respective ranges of dose were: 0-68 rad (#2), 0-96 rad (#5), 0-12 rad (#6), and 0-6 rad (#7).

The pooling of the LSS city-specific samples (Hiroshima and Nagasaki) in the BEIR III (1980) report is an instance of a more fruitful meta-analysis (and data-augmentation). See Fig. 37.

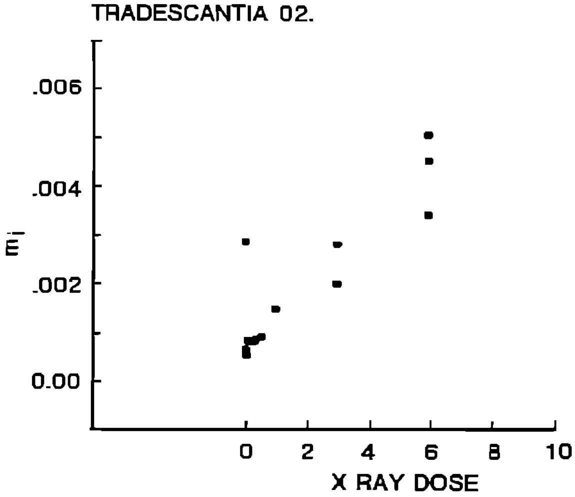


Fig. 1b. Scattergram of the three clusters of observations (m_i, D_i) in Fig. 1a for D < 10 rad. In Fig. 1a the non-uniformity in the distribution of observations is the more apparent; in Fig. 1b the non-uniformity in the replication of observations is the more apparent. Obviously, those levels of dose, D_i, with the greater number of replications will tend to have the greater influence on the goodness-of-fit of the model and data and the estimates of the model parameter vector, $\underline{\beta}$.

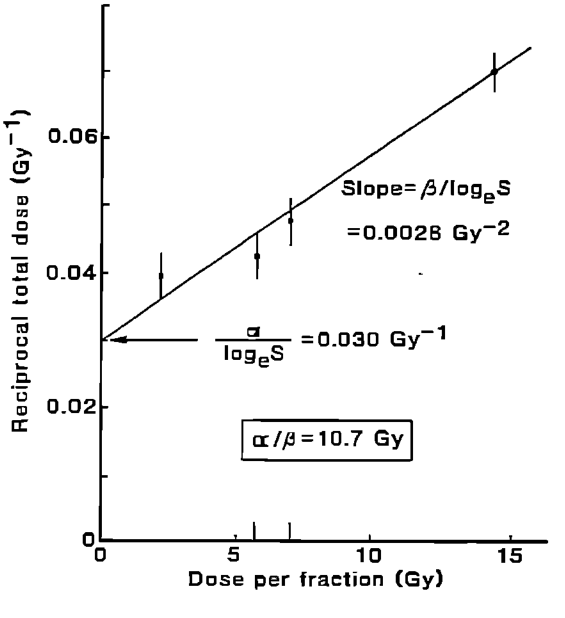


Fig. 2a. Scattergram of the observations (D_i^{-1}, D_i/N_i), $1 \le i \le n = 4$ in a F_e-plot constructed to estimate α/β for the multifraction LQ model. (Reproduced by permission from Fowler, 1984). The equation of the F_e-plot is $D^{-1}(\pi_i) = (\alpha/\log S_i) + (\beta/\log S_i)D(\pi_i)/N_i$, where S_i is the unknown level of survival in an unidentifiable cell population that corresponds to the observed response π_i, where $0 \le \pi_i \le 1$. The F_e-plot is an iso-effect curve; typically, $\pi_i = 0.50$, $1 \le i \le n$. Note the extreme non-uniformity in the distribution of the observations over the levels of D_i/N_i: 75% of the observations are included within the lowest 50% of the D_i/N_i! Note as well that the most extreme observation lies at 15 Gy, well beyond the stipulated range of validity of the multifraction LQ model which Fowler (1984) reports to be $0 \le D_i/N_i \le 10$ Gy. But this observation dominates the sample estimate of α/β: Deletion of this observation inflates the sample estimate of α/β by a factor of two.

As well as dominating the point estimate of α/β the observation at N=1 also controls the precision of the parameter estimates, α/lnS and β/lnS, and the measures of goodness-of-fit of the LQ regression model of these data: Deletion of the observation at N=1 deflates the precision of estimate by a factor of 0.36 and deflates the (adjusted R^2) measure of concordance by a factor of 0.71.

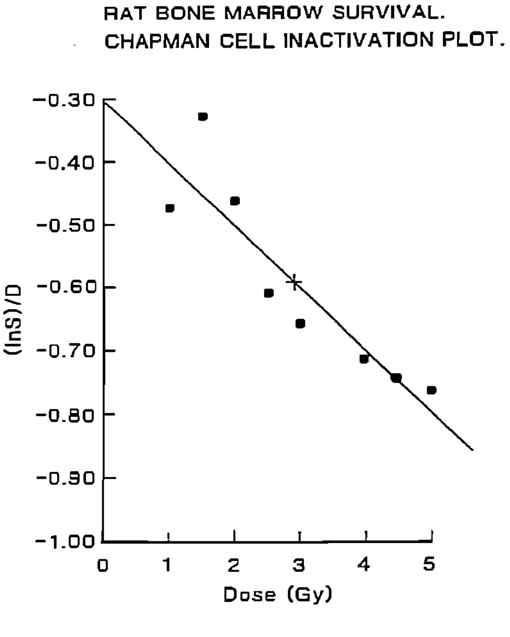


Fig. 2b. Scattergram of the Chapman Cell Inactivation Plot (CIP) of the LQ model of cell survival data. The equation of the plot is $\ln S_i/D_i = \beta_1 + \beta_2 D_i$, $1 \le i \le n$. Note that the equation of the CIP plot can be rewritten in the form of the F_e-plot: $D_i^{-1} = (\alpha/\log S_i) + (\beta_i/\log S_i)D_i$. However, the data of the CIP plot are dose-response observations, whereas, the data of the F_e-plot are iso-effect data. Moreover, for the CIP data the level(s) of survival S_i as well as the identity of the cell population at risk are both known, which, of course, is often not the case for the data of the F_e-plot. For the CIP, "The slope of each plot is the quadratic inactivation constant (β) and the intercept with the zero dose axis is the linear inactivation constant (α)". T. Alper (1980). For the data in the Fig. (Frome and Beauchamp, 1968) the graphical estimates are α = -0.30 Gy^{-1}, β = -0.10 Gy^{-2}, α/β = 3.0 Gy. The straight line is a plot of the line that corresponds to the ordinary least squares estimates of $\alpha(=\beta_1)$ and $\beta(=\beta_3)$: $\bar{\beta}_1 = -0.304$, $\bar{\beta}_2 = -0.098$, $\bar{\beta}_1/\bar{\beta}_2 = 3.112$. It can be shown (See Annex IV, part 6) that the ordinary least squares estimates of α and β in the Chapman equation above correspond to weighted least squares estimates of the parameters of the equation, $\log m_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$, where $w_1 = 10^6$, $w_i = D_i^{-2}$, $2 \le i \le n = 9$. This, of course, means that the observation at D = 0 dominates the sample estimates of α, β, and α/β. Thus, the sample estimate of α/β is dominated by the observation at largest value of $D(\pi_i)/N_i$ in the F_e-plot and at the smallest value of D_i in the CIP plot.

For the data of Fig. 2a the marked non-uniformity in the respective influences of the several observations in the sample is due to the flawed experimental design procedure, whereas for the data of Fig. 2b it is due to the flawed data analysis procedures.

The correct estimates of α, β, and α/β are obtained from the Poisson regression model of the LQ hypothesis for these data, $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$, where m_i is the Poisson rate constant. The estimate of the parameter vector is obtained by iterative reweighted least squares (IRLS) methods: $\bar{\beta}_1 = -0.436$ Gy^{-1}, $\bar{\beta}_2 = -0.074$ Gy^{-2}, and $\bar{\beta}_1/\bar{\beta}_2 = 5.883$. The estimate of α/β obtained from the CIP is only 50% of the correct value.

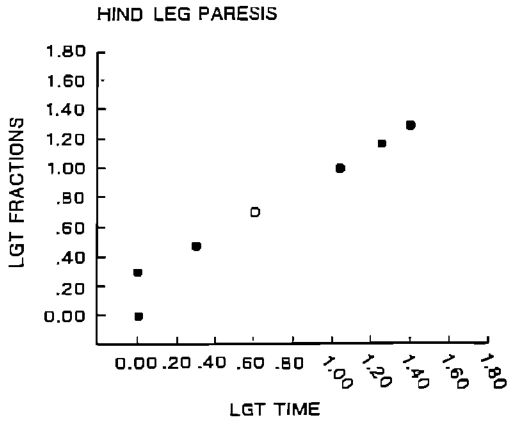


Fig. 3. Scatterplot of the distribution of observations in the dose-response experiment described in Tucker and Thames (1983). In this experiment the end-point was hind-leg paresis in rats and the response was quantal: $0 \leq r_i \leq n_i$, where r_i is the number of responders in n_i at risk at (D_i, N_i, T_i), $1 \leq i \leq n$. There were between 2 and 4 different levels of dose, D_i, at each of seven levels of fractions and time (N_i, T_i), $1 \leq i \leq 19$. The closed symbols identify those levels of (N_i, T_i) at which the response was not extreme - neither $r_i = 0$ nor $r_i = n_i$ - at one or more levels of D_i, where r_i is the number of responders out of n_i at risk at (D_i, N_i, T_i). The plot discloses the presence of extreme multicollinearity in the sample, that is, N_i and T_i are highly correlated. The degree of multicollinearity in this experiment is nearly as high as in the non-experimental data of Fig. 4, suggesting that the experimental design has been poorly chosen since it is known apriori that the response at a given level of total dose D_i may be modulated by the levels of N_i and T_i.

Multicollinear data are weak data from which to construct regression models since the presence of high correlations between the predictor variables inflates both the parameter estimates, $\bar{\beta}_j$ and their standard errors, $\sqrt{Var(\bar{\beta}_j)}$. Moreover, it is often the case that some of the parameter estimates, $\hat{\beta}_j$, may have the wrong sign, i.e., signs that are inconsistent with apriori information on the response at issue.

Experimental data in which there is multicollinearity can sometimes be salvaged by data augmentation. However, it is better to choose orthogonal distributions of observations in the design of the experiment. Examples of orthogonal designs are presented in Figs. 34b and 34c.

Non-experimental data in which two or more predictor variables are highly correlated - as in Fig. 4 - can sometimes be salvaged by data augmentation or by the Bayesian regression maneuvers of Mixed estimation or Ridge regression. See part 7 Statistical Methods. II and Figs. 35 and 38 of this syllabus.

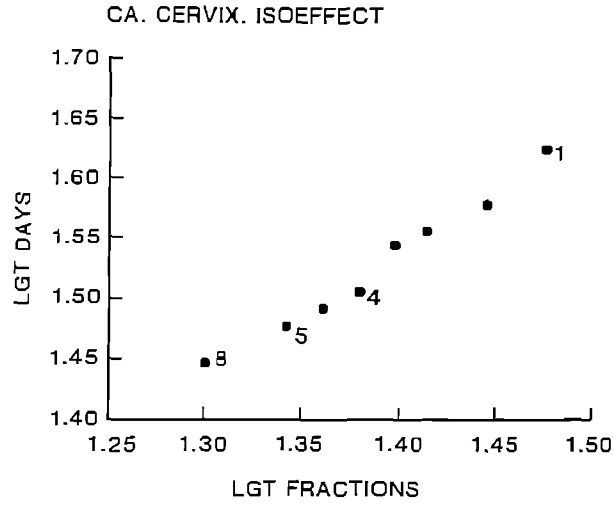


Fig. 4a. Scatterplot of the distribution of a set of non-experimental clinical data. (Supe et al, 1983). In these data the end-point was tumor ablation and the response was quantal, $0 < r_i/n_i < 1.0$, $1 \leq i \leq 8$, where r_i is the number of responders out of n_i at risk at (D_i, N_i, T_i). The data are grouped: $n_i > 1$. The numbers refer to the index number of the observation, that is, the row number in the observation matrix [y, X].

For these data the common level of response - the "effect" in isoeffect - was $\pi = 0.55$. A Pearson chi-squared test of the null hypothesis that $\pi_1 = \_ = \pi_8 = 0.55$ gives a p-value in the extreme lower tail of the chi-squared distribution - suggesting that the "fit" is "too good" to have arisen by random sampling from a population for which $\pi = 0.55$. (See Jeffreys, 1961.)

This is an instance of pooled data in which the data are "too homogeneous". The Tradescantia data of Fig. 1 provide an instance in which the pooled data are "too heterogeneous": It can be shown that these latter data comprise four mutually inconsistent experiments - on the evidence of the respective LQ models.

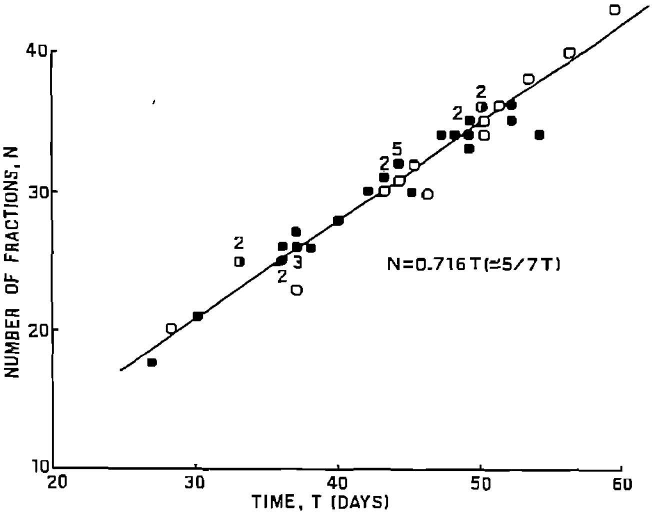


Fig. 4b. Scatterplot of the distribution of observations on control (ablation) of sq. ca. oropharynx in radiotherapy patients at 3 years post-treatment in a set of non-experimental clinical data (Herbert, 1986a). In these data, the end-point was tumor ablation and the response was quantal, $0 \leq r_i/n_i \leq 1.0$, $1 \leq i \leq n$. The data are ungrouped: $n_i = 1$. Thus, $r_i = 0$ or $r_i = n_i = 1$, $1 \leq i \leq n = 46$, where there were r_i responders out of n_i at risk at (D_i, N_i, T_i). The numbers refer to number of observations at a given level of (N_i, T_i). The N_i and T_i are highly correlated, $\rho(N,T) = 0.980$, owing to the "standard of practice" constraint: Treatments are given daily except for Saturday and Sunday: $N_i = 0.716T_i$ ($\simeq 5T_i/7$). The filled symbols identify the responders, the open symbols the non-responders.

of N and T, the amount of information supplied by each experiment was different. See Annex II part 3.

The presence of multicollinearity signifies that one or more treatment variables are strong linear functions of another variable that is included in the sample; that is, the set of variables are highly correlated. For example, the correlation coefficient of N and T may exceed 0.97 in the distributions of observations that may be found in both experimental and non-experimental (clinical) studies. See Annexes II, part 3, III, part 4, IV, part 3. The distributions from the studies examined in Annexes II, part 3, and IV, part 3 are presented in Figs. 3 and 4, respectively. It is the received practice to construct isoeffect equations from highly correlated clinical data such as shown in Fig. 4. For example, Supe et al (1983) constructed an isoeffect model from the data of Fig. 4a: $\log D(\pi) = \alpha_0 + \alpha_1 \log N + \alpha_2 \log T$ where $\alpha_0 \equiv$ TSD, $\alpha_1 = 0.18$, $\alpha_2 = 0.06$, and $\pi = 0.55$. Concerning the construction of regression models from such highly correlated data, Davies et al, 1961 remark that, "When observed values of two independent variables are highly correlated, it may well happen that neither regression coefficients tested against its own standard error, is significantly different from zero, but that the regression on both variables accounts for a significant part of the total variation of the dependent variable. This is equivalent to saying that the appropriate confidence ellipse for joint estimation of both coefficients cuts both axes, but does not include the point (0, 0), so that we may not reasonably conclude that both true regression coefficients are simultaneously zero. We would not, ... be justified in omitting both variables on the basis of their unadjusted regression coefficients and standard errors." (Note that the 0.95 confidence ellipse on the parameter estimates for the isoeffect curve constructed from the data of Fig. 4a, $x_1(\pi) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$, where $x_1 = \log D(\pi)$, $x_2 = \log N$, and $x_3 = \log T$, and $\pi = 0.55$, "cuts both axes" in Fig. 36a.)

For highly correlated predictor variables, the values of the respective weights (coefficients) to be assigned to each predictor variable (which describe the size of its effect on the response variable) cannot be determined from the sample data - the data are too weak - but must be determined from non-sample information (See sections 7.2 and 7.11) Another received practice is "variable selection" - simply delete one variable, either N or T, from the model. Formerly it was N that was deleted as, for example, in the early "cube-root law" ($D \propto kT^{0.33}$); more recently it has been T (see Fowler, 1984, 1989). This practice reduces the model to the simple form, $y = a + bx$ in either case. (See section 7.5.)

The effects of the presence of these two flaws - non-uniformity and multi-collinearity - in the distribution of observations on the estimates and inferences to be made from a regression model of $[\underline{y}, X]$ are as follows: For non-uniformly distributed data, a few observations (or even one) will dominate the estimates and inferences made from most regression models of the sample. For models of data in which the distribution of observations has a high level of multicollinearity, both the parameter estimates, say $\hat{\underline{\beta}}$, and the variance of the parameter estimates, $Var(\bar{\underline{\beta}})$, are inflated. And it is often the case that one or more parameter estimates, $\hat{\beta}_j$, with the "wrong" sign will be obtained as well. It is also the case that these two flaws, non-uniformity and multicollinearity, interact in samples in which both are present, to exacerbate the effects of each on the sample estimate of $\underline{\beta}$. See Statistical Methods II in this report. Thus, the bias and precision of estimates and inferences made from regression models of non-uniform and/or multicollinear data are profoundly affected. The evidential value of such estimates and inferences - and thus their "believability" - is degraded thereby.

6.1.4 The amount of information in an observation. Weights.

As is well-known, the amount of information on dose-response that the observation at the i^{th} level of treatment, $\underline{x}_i^T$, provides to the investigator, that is, its weight, w_i, depends both on the number of subjects (cells, tissues, animals, etc.) at risk at $\underline{x}_i^T$ and on the form of the distribution of the random part of the response, ε_i, that is, whether the distribution is Normal, Binomial, Poisson, etc. However, our secondary analyses disclosed that in most of the studies reviewed, the respective weights of the sample observations were either ignored or arbitrarily - albeit unwittingly - inflated or deflated to extreme levels by the current modelling practices. We defer consideration of the arbitrary adjustment of weights to section 6.2 and consider here only the effects of misspecification of the distribution of the random part of the model.

It will be recalled that for the Normal distribution the expectation and variance of the distribution are independent, whereas for the Poisson and Binomial distributions this is not the case. Therefore, one immediate consequence of misspecification of the random part of the response is that the respective inherent weights, w_i, $1 \leq i \leq n$, of each of the n observations of the sample are incorrect if the distribution of the random part of the response is nonNormal and the slope of the dose variable is non-zero. Incorrectly weighted observations influence both the measures of goodness-of-fit of the model and the bias and variance of the sample estimates of the parameters of the model and functions thereof, such as the predicted response. However, it is the practice of most investigators to ignore the dimension, or metric, of the observed response (counts, proportions, etc.) and to treat all observed responses, y_i, as continuous measurements, $-\infty < y_i < \infty$, with a (conditionally) Normal distribution implied.

The widespread practice of investigators to ignore the sample evidence that the non-Normally distributed responses in their observations must have inherently different variances (since the variance of the conditional distribution of response is a function of the expected value for Binomial and Poisson responses) - and hence weights - might well explain, as a result of this ontological weakness, two recurring flaws that we found in the experimental designs of the studies we examined. Both devolve from the contingently different variances of the observed responses in those experiments in which the response has a Binomial distribution.

The first flaw is that in the studies we reviewed the numbers at risk specified by the design, n_i at x_i^T, $1 \leq i \leq n$, are typically much too small, e.g., n_i is typically 5 or 6, but $25 \leq n_i \leq 35$ is required by good design practice. See Annex II, part 3 for examples and discussion. The important principle of parsimony in the construction of a regression model of a set of observations - frequently justified, especially by physicists, on Mach's principle of economy of thought (vide supra) - appears to have been transmogrified, in practice, into a principle of parsimony in the acquisition of the observations themselves, since "economy of thought" is also achieved by economizing on the number of observations, as well as the number of parameters, that one has to think about and "explain" - in the context of a received model. See the Shellabarger, et al, experiments on mammary neoplasia for an example of an application of the principle of "parsimony in observation" as discussed above in section 5.1.2d.

As an important collateral perspective on this practice of parsimony in observation, note that it is also the case that most of the currently designed clinical therapeutic trials (randomized controlled trial, RCT) have sample sizes that are much smaller than is recommended for statistically adequate designs. For example, Klein et al, 1986 found that in a survey of 28 published RCTs, "The average sample size per treatment arm of these studies was 21 and only 2 of the 28 trials had more than 50 patients per treatment arm." And, "The most depressing feature about clinical research in many disease areas is that investigators and pharmaceutical companies continue to put their effort into small single-centre studies based on 20 patients or less." S. Pocock and J. Hughes (1990). But Zelen (1982) recommends that, "Comparative trials should be planned with a minimum of 100-200 patients per treatment," and notes that "Trials with fewer patients are likely to produce more false-positive results than true positive results." Zelen (1982) remarks further that, "We may now have reached an impasse in cancer chemotherapy in which there are so many false-positive therapies in the clinic that they are overwhelming the true positives. Many physicians have a deep feeling that very few patients may be benefiting in any significant way from most of the treatments being administered."

In the context of randomized controlled trials small sample sizes result in a loss of statistical power also and hence in an increase in the probability of false negative, as well as false positive, reports. As a consequence treatments that produce a modest, but worthwhile, difference in response from either a medical or public health perspective may well be overlooked. The newer statistical discipline, meta-analysis, has been largely motivated and developed as a method for the salvage of these false negative clinical trials. Meta-analysis, it will be recalled, describes a set of methods for combining, or pooling, the data, or, more often, the results, which are usually described by the summary statistics - e.g., measures of concordance or association such as chi-squared or estimates of parameters such as slopes of regression lines - of several presumably related studies, or experiments, to come to a single overall conclusion. Sacks et al, 1987 notes that, "A recent review of meta-analysis in the field of public health has emphasized its growing importance." See Louis et al, 1985. But, as DuMouchel (1989) has remarked, "Critical judgments about the accuracy, quality and methodologies of the individual studies are necessary in order to combine them intelligently." We remark that secondary analyses of the individual studies, such as we have described in Annexes II-IV, are necessary to achieve the insights and perspectives required for these critical judgments that must precede any integration, or synthesis, of a set of studies.

As noted above, small numbers of patients recruited to a clinical trial also lead to an increased probability of false positive results which are more likely to be published simply because they are positive - the so-called publication bias. Therefore, Zelen (1982) recommends replication of any study as the appropriate salvage - or protective - maneuver against accepting false positive results: "... all positive results should be independently confirmed. This will lower the false positive rate and raise the true positive rate. Physicians in practice should exercise caution in adopting a new therapy if there is no independent confirmation." We shall recur to the epistemological requirement for replication again below.

The second recurring weakness that we found in experimental designs for models of non-Normal responses is that the number at risk, n_i, at a given level of treatment variables and covariates, $\underline{x}_i^T$, often varies quite widely within the set of dose-response experiments reported in a given study, giving rise to a second sort of quite troublesome non-uniformity. For example, in a paper on radiation toxicity in which the response was quantal and hence the distribution of response was Binomial, published in 1984, we find that $3 \leq n_i \leq 12$, where n_i is the number at risk at $\underline{x}_i^T$, $1 \leq i \leq n$. Obviously, the estimates and inferences obtained from those studies in which the numbers at risk vary widely, will tend to be dominated by those observations at the larger n_i - since, other things being equal, those observations will tend to have the greater weights. Thus, the study results will be biased if non-uniformity in either the distribution of the numbers at risk, n_i, or in the distribution of the levels of treatment, $\underline{x}_i^T$, is present. ("Generally speaking, binomial observations based on larger numbers of cases and controls have greater influence than those based on smaller numbers. The same may be true for outlying observations, i.e., those for which exposure values $\underline{x}$ are far removed from the rest of the data" Breslow and Storer, 1985.)

We also encountered non-uniform replication of observations in an experiment in which the response has a Poisson distribution: radiation mutagenesis in Tradescantia (Sparrow et al 1972). These are the data to which a single fraction LQ model was fitted in NCRP 64 (1980). They consisted of a pooled set of four dose-response experiments, each with five levels of dose covering different ranges: 0-6 rad; 0-12 rad; 0-68 rad; 0-96 rad. Although several of the dose levels in one experiment were repeated in one or more of the other experiments, the replication of dose levels achieved in the pooled data was markedly non-uniform. The degree of this kind of non-uniformity is well shown in Figs. 1a and 1b. Within each of the four experiments the doses increased geometrically rather than arithmetically - giving rise to a non-uniformity of the first kind. The numbers (of stamen hairs) at risk varied markedly between the four experiments - by over a factor of 10. As a consequence, the sample estimate of $\underline{\beta}$, the parameter vector of the LQ model obtained from the pooled data, was dominated by the information included in the data of those experiments, or studies (of five dose levels each), that had the largest numbers at risk. This gives rise to a version of the Simpson's paradox that sometimes biases the conclusions obtained from a meta-analysis, or integration, of a set of studies. Halvorsen's comments anent another meta- analysis are appropriate: "The biggest problem is that we may be introducing an analogue of Simpson's paradox; when true effects vary from study to study, the size and even the direction of the combined results may depend heavily on such extraneous features as which studios were largest and hence tend to dominate the analysis" (Halvorsen, 1986).

The failure of the received practice to correctly identify the nature (Binomial, Normal, Poisson, etc.) of the distribution of the random part of observed response in the construction of dose-response models can lead to still other ontological weaknesses. In the case of quantal radiation responses, which have a Binomial distribution, the weaknesses have to do with the shape of the dose-response curve, or surface, and the shape of the concomitant tolerance distribution. The Binomial parameter, π_i, of the (conditional) distribution of response has a finite range: $0 \leq \pi_i \leq 1$. In order to use regression methods a transformation of the response, say probit or logit, is required: $\pi_i \longrightarrow z_i$ where z_i is the probit or logit transform and $-\infty < z_i < \infty$. This transformation gives a sigmoid dose-response curve that implies both a unimodal distribution of tolerance - usually logNormal - and the presence of an effective non-zero "threshold" dose in the dose-response curve. On the other hand, the frequently deployed linear model of a Binomial response, say $\pi = \alpha_0 + \alpha_1 D$, implies a rectangular, or uniform, distribution of tolerance dose and the absence of a non-zero threshold. Both of these implied features are unlikely, a priori; a rectangular tolerance distribution is quite implausible and threshold dose-response curves are common to most noxious agents. The probit model gives a biphasic dose-response curve, the linear model gives a monophasic dose-response curve. (The respective models of quantal responses are compared in Annex III, part 6.

We have suggested above that the form of the (conditional) distribution of the random part of the response, Normal, Binomial, or Poisson, can be determined, a priori, from the metric of the observations which may be a (continuous) measurement, a proportion, or a count, respectively. This is generally true. However, for some samples of quantal data the conditional distribution of response may be Beta Binomial rather than Binomial and for some samples of count data, the distribution may be Negative Binomial rather than Poisson. In these cases it may be necessary to determine the precise form of the (conditional) distribution of the random part, ε_i, of the response, y_i, as well as the form of the deterministic part, μ_i, from the sample itself. The LQ model of the Tradescantia data that is described in NCRP 64 presents such a case. These are the data described in Figs. 1a and 1b. Some of the results of our secondary analyses of these data are briefly described in Figs. 5-8 and a more complete discussion can be found in Annex III, part 5. The problem with the regression analysis of the Poisson data of Fig. 1 that is presented in NCRP 64 (1980) is quite analogous to that just described for the linear model, $\pi = \alpha_0 + \alpha_1 D$, of quantal data.

The form of the LQ model of the Tradescantia data selected by NCRP 64 is $\log I = \log(\alpha D + \beta D^2)$ where I is the excess incidence and D is the radiation dose. There is more than one weakness in this formulation, but we address here only the misspecification of the (conditional) distribution random part of the response: The above form of the LQ model implies that the distribution of the random part of the response is Negative Binomial rather than Poisson. In Annex III, part 5, we show that the above formulation is a specific case of the so-called power-Normal model of Snee (1986) which for the case at issue may be written in the general form

$$I^\lambda = [\alpha D + \beta D^2]^\lambda$$

where λ is a parameter to be estimated from the data. The value obtained for λ depends on the form of the distribution of the random part of the response: $\hat{\lambda} = 0$ identifies the Negative Binomial distribution; $\hat{\lambda} = 0.5$ identifies the Poisson distribution. If λ is correctly chosen then the residuals, e_i, for the model have a Normal distribution - if the deterministic part of the model is also correctly specified.

For $\lambda \longrightarrow 0$ the power Normal model takes the form

$$\log I = \log[\alpha D + \beta D^2],$$

the form assumed, a priori, in NCRP 64 (1980) and which entails the assumption of a Negative Binomial distribution of the random part of the response. The analyses summarized below in Figs. 5 and 6 show quite vividly that the sample data of Fig. 1 are not consistent with this a priori assumption. Figures 7 and 8 describe an alternative analysis of the data of Fig. 1.

Failure to correctly identify the form of the random part of a Poisson-type response, results in incorrect estimates of the variance of the parameter estimates and, if the distribution of observations is non-uniform - as is the case of the data of Fig. 1 - in incorrect estimates of the

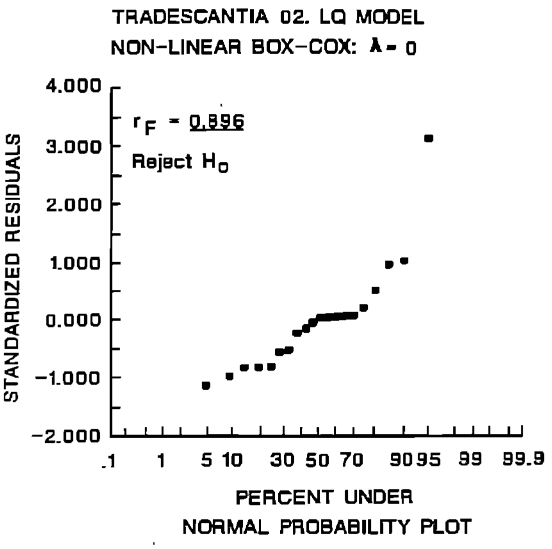


Fig. 5a. Normal probability plot of standardized residuals, $e_i^* = e_i/\sqrt{Var(e)}$, for the power-Normal (non-linear Box-Cox) model, $m_i^\lambda = (\beta_0 + \beta_1 D_i + \beta_2 D_i^2)^\lambda + e_i$, $-\infty < \lambda < \infty$, $1 \le i \le 20$, of the Sparrow et al (1972) data on the incidence of pink mutations in the stamen hairs of Tradescantia. m_i is the incidence rate, mutations per stamen hair. λ is a Normalizing parameter. If the choice of λ is consistent with the form of the distribution of the random part of the observed response then the distribution of the residuals for the model of the data will be Normal. In NCRP 64 the value of λ = 0 was selected, a priori, for the LQ model of these data. This value of λ is consistent with a Negative Binomial distribution of the random part of the observed response. For $\lambda = 0$, the working equation of the model takes the form, $\log m_i = \log(\beta_0 + \beta_1 D_i + \beta_2 D_i^2) + e_i$, and estimates of the parameter vector $\underline{\beta}$ are obtained by iterative least squares methods. The estimates of $\alpha(=\beta_1)$, $\beta(=\beta_2)$ and α/β $(= \beta_1/\beta_2)$ that are presented in NCRP 64 Table 5.1 were obtained from the equation $\log I_i = \log(\beta_1 D_i + \beta_2 D_i^2)$ where $I_i = m_i - m_1$ where m_1 is the response rate at $D_1 = 0$. (Note that this is an "absolute risk" model.) However, the transformation $m_i \longrightarrow I_i$ is equivalent to imposing constraint on the estimate of $\underline{\beta}$: $\underline{r} = R\underline{\beta} + \underline{v}$, $E(\underline{v}) = \underline{0}$, $Var(\underline{v}) = \psi = [0]$ where $m_1 = \beta_0 = r$, $R = (1,0,0)$, and $\underline{0}$ and [0] are the null vector and matrix, respectively. But it can be shown that 1) this represents spurious a priori information on the spontaneous incidence rate, 2) it degrades the "fit" of the model to the data and 3) provides estimates of $\underline{\beta}$ which are a) biased and b) of spuriously high precision.

It is evident from the shape of the Normal probability plot, and the value of the Filliben probability plot correlation coefficient, r_F, that the distribution of these residuals is non-Normal, indicating that $\lambda = 0$ is a poor choice for the Normalizing parameter.

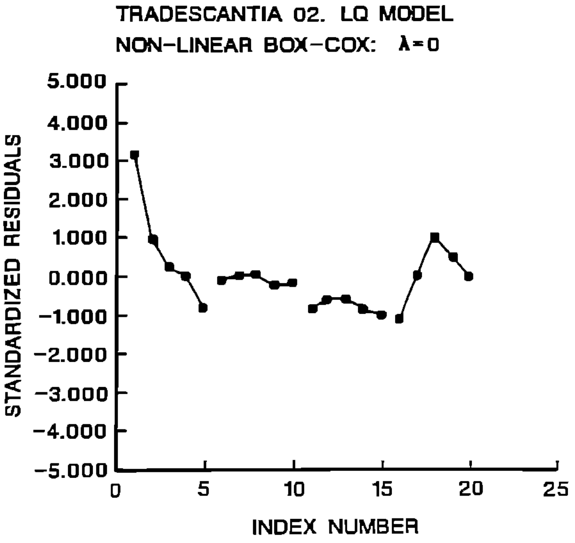


Fig. 5b. Index plot of standardized residuals $e_i^* = e_i/\sqrt{Var(e)}$, for the power-Normal model described in Fig. 5a with $\lambda = 0$. The index numbers, i, $1 \le i \le 20$, are simply the row numbers of the observation matrix, $[\underline{v}, X]$. In the case of the present data the observations have been grouped by experiment: $1 \le i \le 5$ (experiment #5); $6 \le i \le 10$ (experiment #6); $11 \le i \le 15$ (experiment #7); $16 \le i \le 20$ (experiment #2).

Within each experiment the observations are ordered by dose, D_i. It is evident from the plot that the 5 observations in experiment #6 ($0 \le D \le 12$ rad) dominate the estimates of the parameter vector, $\underline{\beta}$, of the LQ model of these data - since these have the smallest standardized residuals, e_i^*. This is to be expected, a priori, since a) the respective average numbers of stamen hairs at risk, n_i, in each experiment are 1926, 52576, 437,317, and 3473 for the experiments numbered 5, 6, 7 and 2 and b) the respective ranges of dose are $0 \le D \le 96$ rads, $0 \le D \le 12$ rads, $0 \le D \le 6$ rad, and $0 \le D \le 68$ rads.

It is evident that the observation in experiment #5 at D = 0 may be a "contaminant".

The joint 0.95 confidence limits on $\bar{\beta}_2$ include 0. Since the non-linear Box-Cox methods typically under-estimate $Var(\bar{\beta}_j)$ this suggests that the LQ model over-fits these data - on the a priori assumption of a (conditional) Negative Binomial distribution for the response.

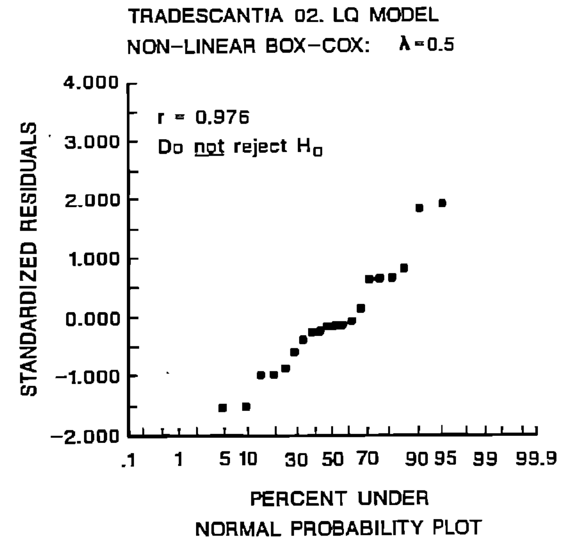


Fig. 6a. Normal probability plot of standardized residuals, $e_i^* = e_i/\sqrt{Var(e)}$, for the power Normal (non-linear Box-Cox) model $m_i^{\lambda} = (\beta_0 + \beta_1 D_i + \beta_2 D_i^2) + e_i$, $1 \leq i \leq 20$, of the Sparrow et al (1972) mutagenesis data of Fig. 5 for $\lambda = 0.5$. This estimate of λ was obtained from the data by a Box and Cox procedure, and is consistent with a Poisson distribution for the random part of the observed response. It is evident from the plot and the Filliben probability plot correlation coefficient, r_F, that the choice, $\lambda = 0.5$ for the Normalizing parameter is appropriate. (See Fig. 6c.)

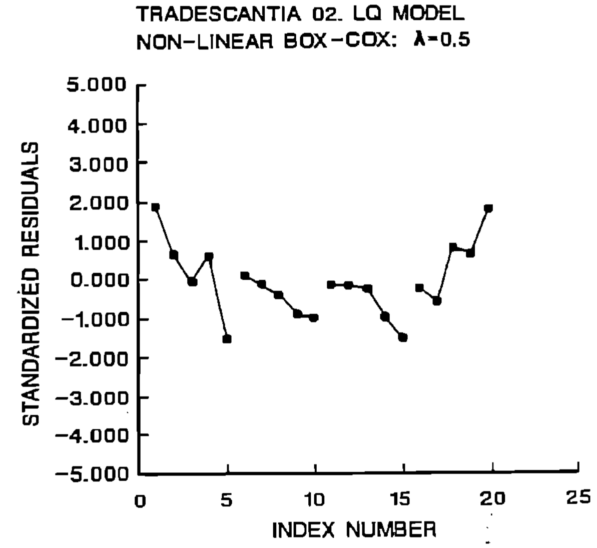


Fig. 6b. Index plot of standardized residuals $e_i^* = e_i/\sqrt{Var(e)}$ for the power-Normal model described in Fig. 6a with $\lambda = 0.5$. As in the case of Fig. 5b it is evident that the 5 observations in experiment #6 ($0 \leq D \leq 12$ rad) dominate the estimates of the parameter vector $\underline{\beta}$ of the LQ model of these data. Although on the evidence of both the aggregate and case statistics the LQ linear predictor, η_i, "fits" these data on the sample estimate $\lambda = 0.52$, the joint 0.95 confidence limits on β_0 include zero. However, it is well-known that the spontaneous ($D = 0$) mutation rate may be quite large. This suggests that the LQ model does not intersect these data (As Robins and Greenland (1986) remark, it is a good strategy "... to choose a model that is consistent with the data and yields parameter estimates consistent with prior beliefs.")

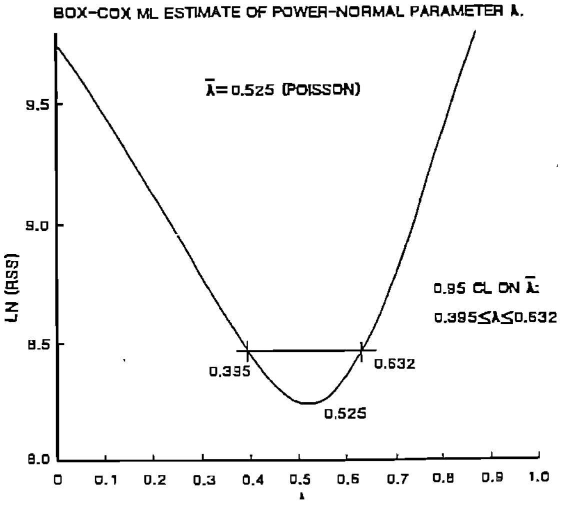


Fig. 6c. Plot of log of residual sum of squares vs Normalizing parameter λ for the power-Normal model $m_{i\lambda} = (\beta_0 + \beta_1 D_i + \beta_2 D_i^2)^{\lambda} + e_i$ of the Tradescantia 02 data in Fig. 1a. It is obvious that for those data $\lambda \simeq 0.5$, suggesting that the random part of the observed response has a Poisson distribution, rather than the Negative Binomial distribution ($\lambda = 0$) assumed in the NCRP 64 (1980) analyses.

This Box-Cox technique turns the problem of identification (of the form of the distribution of e_i) into a problem of estimation (of λ). This same principle can be employed in discrimination between the forms of the deterministic parts for rival non-nested models. For example, to discriminate between the Target theory, say $f(D_i)$, and LQ, say $g(D_i)$, models of cell-survival, both can be embedded in a larger model that includes each as a special case, indexed by the parameter λ. That is, $m_i = f(D_i)^{\lambda} * g(D_i)^{(1-\lambda)}$. Point and interval estimates of λ can be used to select either $f(D_i)$ for which $\lambda = 1$, or $g(D_i)$, for which $\lambda = 0$. But although this can be done in principle, it usually cannot be done in practice since it would require a minimum of $3 + 3 + 6 = 12$ observations, (m_i, D_i), $1 \leq i \leq 12$ (more would be preferable).

parameter vector as well. N.B. It may be recalled that for the Binomial, Negative Binomial, and Poisson distributions the first and second moments are related as follows:

Binomial	variance < mean
Poisson	variance = mean
Negative Binomial	variance > mean

It is of interest to note that the alternative analysis of these mutagenesis data disclosed that the estimates of β for the LQ model that were obtained from each of the four experiments in the pool were study-specific and hence mutually incompatible. Both a study effect and a study times dose interaction effect could be identified quite unambiguously and were included in an expanded regression model using indicator variables (vide infra). It is important to emphasize that if the three indicator variables that identified the four studies that were pooled are omitted from the model, the Poisson linear LQ model does not fit the Tradescantia mutagenesis data on the evidence of statistically adequate measures and that for the expanded model the coefficients of the terms for the indicator variables and their interactions differed significantly from zero. That is to say, the LQ model of the Tradescantia data that is presented in NCRP 64 Table 5.1 is underfit.

Important Topics:

T. Kuhn, C. Peirce, K. Popper, H. Jeffreys, F. Bacon, J. Ziman, J. Nisbett, L. Ross, K. Tversky, T. Kahneman. Belief perseverance. Null hypothesis. F_e-plot. Affirming the consequent. Prejudice against the null hypothesis. Positive spin in scientific papers. Ontological weaknesses. Stochastic and deterministic parts of response. Multivariate response (two senses). Joint and marginal probabilities. Overfitting. Underfitting. Weight of an observation. Parsimony. Replication. Link function. Probability distribution.

6.2 Epistemological weaknesses in received modelling praxis

"The central problem of epistemology has always been and still is the problem of the growth of knowledge."

K. Popper, 1959

"Much of what is published goes unchallenged, may be untrue, and probably nobody knows. Does anybody care? Do the methods used to obtain results matter anymore?"

A. Neufeld, 1986

6.2.1 Study data, some aspects of

The secondary analyses described in Annexes II, III and IV also disclosed the presence of several epistemological weaknesses in the regression models of radiation dose-response that are described in the peer-reviewed literature. These weaknesses are discussed in this section. However, a few general remarks anent data acquisition and analysis must be made as a preliminary.

Because of the markedly quantitative nature of the validated reviews required for this report the Task Group 1 has been greatly concerned with both the adequacy and availability of the primary data of the studies evaluated. But here it must be remarked that in order to evaluate the methodology which yielded the results presented in two of these studies (Annex II, parts 3 and 4, a review of Tucker and Thames, 1983; Appendix I, reviews of von Essen, 1960) we have found it necessary to interpret the term "data" in the larger sense as given in Webster's Third New International Dictionary, 1966 edition: "datum. 1a: something that is given either from being experientially encountered or from being admitted or assumed for specific purposes: a fact or principle granted or presented: something upon which an inference or an argument is based or from which an intellectual system of any sort is constructed." For although the published data in these two studies do not, in the event, arise from an "experiential encounter" of the investigator, - they are, in fact, fictitious; that is, they are "non-experiential" - evidently they do describe something

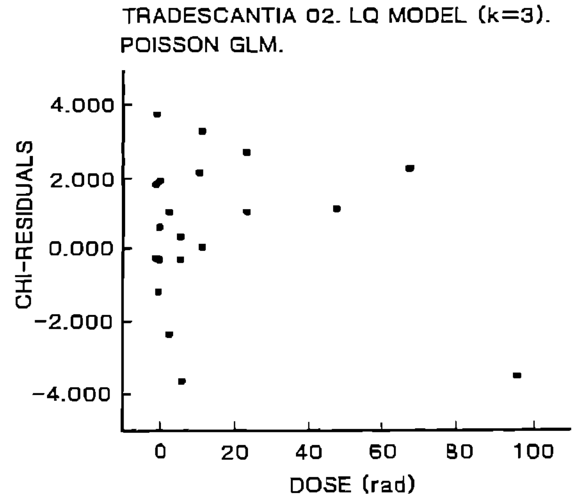


Fig. 7a. Plot of chi-residuals, $\chi_i = (y_i - \hat{m}_i c_i)/\sqrt{\hat{m}_i c_i}$, vs the dose for the Poisson linear model with linear predictor $\eta_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$. $1 \leq i \leq n = 20$. It is obvious that the model and data are not concordant, although there is no strong pattern evident. The data are clearly "over-dispersed". This is consistent with the evidence of the aggregate statistic RSS $= \chi_c^2 = \Sigma\chi_i^2 = 86.3$ which is distributed as Pearson chi-squared on 17 degrees of freedom. For this value of χ_c^2, $P(\chi^2 > \chi_c^2 \beta) = 2{*}10^{-10}$.

We note that the addition of cubic and quartic terms in dose, D_i^3 and D_i^4, respectively, decreased $\Sigma\chi_i^2$ by decrements that were statistically significant. However, the data rejected both of these augmented models on the basis of the respective overall measures of fit, $\Sigma\chi_i^2$, evidence that although one of two models may "fit better" than the other, it still may "not fit" the data. It is always necessary to check whether the "better-fitting" of two models, actually "fits" the data - on a statistically adequate measure.

Note that although the BEIR III report cites the Tradescantia mutagenesis data as empirical support for the LQ model of low LET radiation effects, the respective parameterizations of the LQ hypothesis differ between the models of mutagenesis (NCRP 64, 1980) and the BEIR III LSS data (1980). The residuals plot in this figure refers to the LQ model of the mutagenesis data in which the BEIR III parameterization is deployed. As Darby (1986) has remarked, "As one step in the process of establishing whether it is appropriate to generalize from the experience of [one set of] studies _ to other populations exposed to radiation, it is useful to analyze data from the two [or more] studies using so far as is possible, identical methodology; and then to compare the findings. In this way any differences between the two studies will be highlighted, and if none are found the results will be in suitable form for combination."

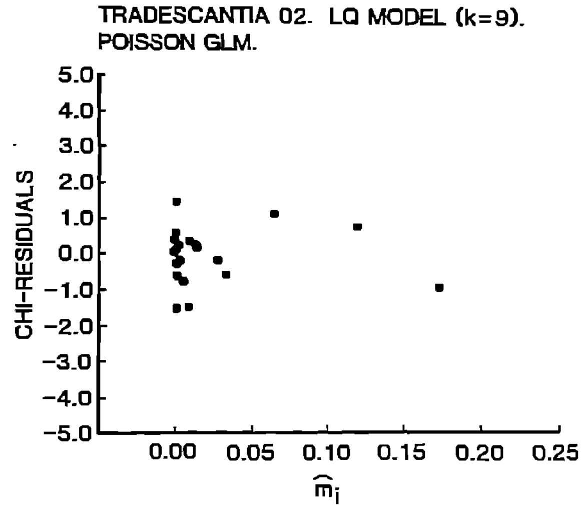


Fig. 7b. Plot of chi-residuals $\chi_i = (y_i - \hat{m}_i c_i)/\sqrt{\hat{m}_i c_i}$ vs the predicted value, m_i, for the Poisson linear model with linear predictor $\eta_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 E_1 + \beta_4 E_2 + \beta_5 E_3 + \beta_6 E_1 D_i + \beta_7 E_2 D_i + \beta_8 E_3 D_i$ where the E_j, $1 \leq j \leq 3$ are the indicator variables:

	Experiment			
	#2	#5	#6	#7
E_1	0	0	1	0
E2	0	0	0	1
E3	1	0	0	0

It is obvious that the augmented model, which describes both an experiment effect and an experiment*dose interaction provides an adequate fit to these data. Note also that this augmented model can be more readily interpreted than a model with cubic and quartic terms in dose.

Note that for the parallel model (LQ-L) of the BEIR III (1980) LSS data in which the city-specific (Hiroshima and Nagasaki) data were pooled (See Fig. 35) the coefficients for the dummy variable for city were non-significant, indicating that the hypothesis of homogeneity of the two sets of data could not be rejected.

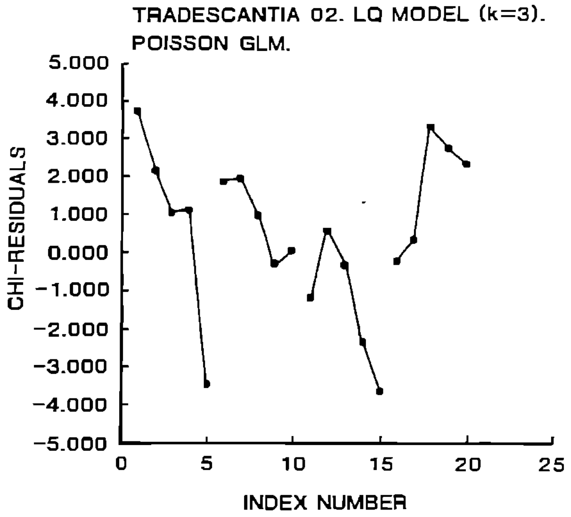


Fig. 8a. Index plot of the chi-residuals for the Poisson linear model of the Sparrow et al (1972) Tradescantia mutagenesis data over the range 0-100 rad X-rays. The linear predictor is that described in Fig. 7a.

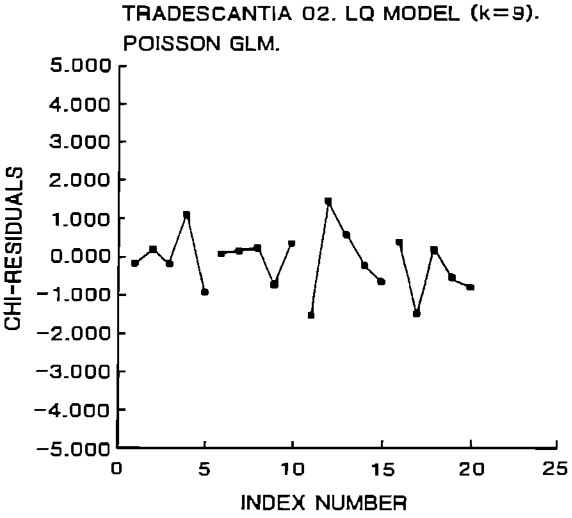


Fig. 8b. Index plot of the chi-residuals for the Poisson linear model of the Sparrow et al (1972) Tradescantia mutagenesis data over the range 0-100 rad X-rays. The linear predictor is that described in Fig. 7b which includes indicator variables to identify the data of four experiments that were pooled.

Box, Hunter, and Hunter (1978) have remarked that, "All the information relating to the possible <u>inadequacy</u> of a tentatively entertained model is contained in the residuals. Plots of the residuals, therefore, can reveal particular aspects of the model that should be improved. Consequently, as a matter of course, residuals should always be plotted in any way that might shed light on pertinent questions."

that was "... admitted, or assumed for specific purposes," and "something upon which an inference or an argument is based ..." Such studies also serve to remind us once more of the lingering presence of the past in current scientific practice: "Within the ancient and medieval tradition, many experiments prove on examination to have been 'thought experiments', the construction of potential experimental situations the outcome of which could safely be foretold ..." T. Kuhn, 1977. In Annex II, part 4 we have discussed the Tucker and Thames 1983 study of radiation toxicity (hind leg paresis in the rat) as a, "thought experiment".

It is the case, of course, that even experiential data may not be adequate to answer the question addressed in a study; that is, to repeat once more Tukey's (1983) remarks, "1) The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. 2) The data may not even contain the appearance of an answer, although we should look for appearances and then report them labeled as such rather than answers." Thus, in those studies in which the data were weak, what we report in Annexes II-IV must be, obviously, regarded as the "appearance of the answer" that would be obtained by the deployment of statistically adequate methods and criteria in cognate data from which such flaws were absent.

It is also useful to recall that, contrary to etymology, data are gotten, not given: "Hence they are eclectic, chosen in a particular context, from a particular viewpoint, and even for a particular purpose. Who gets them, how and why are essentially social questions having determinative effect on accuracy, adequacy and availability." I. Hoos, 1980. (emphasis added)

6.2.2 Some epistemological weaknesses

"Just as there are accepted standards for laboratory procedures, there are mathematical and statistical standards that modelers should adhere to."

S. Moolgavkar, 1991

"Can't anybody out there play this god damn game?"

Leo Durocher, 1953

"False facts are highly injurious to the progress of science for they often endure long; but false views, if supported by some evidence, do little harm, for everyone takes salutory pleasure in proving their falseness."

C. Darwin, 1861

The weakness in the data - non-uniformity, multicollinearity, small numbers, n_i, of animals at risk at each level of treatment, $\underline{x}_i^T$, etc. - remarked in part 6.1, a consequence of ontological weaknesses in received practice, have, of course, their epistemological counter-parts and consequences, some of which we shall examine below in section 7. However, there are still other epistemological weaknesses present in those studies listed in Table 1 that were disclosed to secondary analysis. We now consider these as well:

a) The concordance, or goodness-of-fit, of the specified model and sample data of the study was rarely assessed in the original studies by statistically adequate measures and criteria. In those few studies in which statistically adequate measures of concordance are deployed, the criteria are often misapplied - or ignored. Of course, "... the primary aim of data analysis is to explain or account for the behaviour of the data, not simply to get the best fit." (D. Bates and D. Watts, 1988). But, "... a model that does not fit the data will be of no interest to the modeller even if it may be shown to have various other desirable properties." (D. Ratkowsky, 1983).

b) The concordance, or consistency, of a priori and sample information on the model parameters, say $\underline{\beta}$, or on the response, say $\underline{x}^T\underline{\beta}$, was never explicitly assessed in any study, although such non-sample information was frequently deployed in the studies that are listed in Table 1, either to stabilize the sample estimates of the parameter vector when those estimates, $\hat{\underline{\beta}}$, are imprecise - owing to the weakness of the data - or to fix the level of response at a value selected, a priori, by the investigator, say, in order to reduce the number of parameters in the model to be estimated from the sample data (e.g., "the survival transformation"; see Fig. 12b). That is, in many papers it is never adequately demonstrated that the received models possess the desiderata of accuracy with respect to the sample data and of consistency with respect to the non-sample, or a priori, information on the matter at issue that has been deployed by the investigator. See Annex IV, parts 3, 4, 5, and 6. As Robins and Greenland (1986) have remarked, it is good modelling strategy "... to choose a model that is consistent with the data and yields parameter estimates consistent with prior beliefs." However, the consistency of the model and the estimates of its parameter vector, $\underline{\beta}$, as well as functions thereof, $f(\underline{\beta})$ with the sample data and with cognate a priori information must be demonstrated - not simply assumed, as is currently the received practice - in each case in which the model is deployed.

c) As remarked above, it is often desirable, at times it is even necessary, (Since, "All non-experimental data sets are simply too weak to allow sensible inferences in the absence of supplementary information." Leamer, 1986), to combine with the sample information on the model parameter $\underline{\beta}$, supplemental non-sample information on the level of response, y^*, at a specified level of dose and covariates, $\underline{x}^{*T}$, or on the parameter vector, $\underline{\beta}$, of the model. However, the current practice in radiobiology does not provide the concepts, the methods, and/or the criteria for achieving these useful combinations in either a systematic or a statistically adequate manner.

d) The a priori information (on the size, sign, and statistical significance of the parameters, β_j, $0 \leq j \leq k$, or on the level of response at a specified level of treatment variables, $\underline{x}_i^T$, $1 \leq i \leq n$) that is deployed, for whatever purposes, by the investigator in the construction of the models in the studies we reviewed is, if relevant, often either inherently quite implausible, or quite weak. And, in some studies, the a priori information that is introduced by the investigator is altogether alien to the process that generated the sample data.

The concepts, methods, and criteria for assessing both the concordance of the proposed model with the sample data and the consistency of the sample information with the cognate non-sample information on the parameter vector and/or on the level of response, y^* at a specified level of $\underline{x}^{*T}$ are provided by modern regression analysis. Measures and criteria based on regression diagnostics, or case statistics, such as, the sizes of the individual residuals, $e_i = y_i - \hat{y}_i$, of a model supplement the classical tests of hypotheses on concordance that are based on aggregate statistics, such as the sum of squared residuals, Σe_i^2. Bayesian regression procedures provide formal, systematic, statistically adequate, intuitively satisfying, and generally useful methods both for combining, or "mixing", both conceptual and empirical evidence on the regression model itself, on the model parameter, $\underline{\beta}$, or on a linear function of $\underline{\beta}$, say the response, $\underline{x}^T\underline{\beta}$ at $\underline{x}^T$ and for assessing the validity and utility of the combination achieved thereby. More generally, these procedures provide both methods and criteria for combining sample with non-sample information on issues of interest. These procedures are especially useful, of course, when the sample evidence is weak - that is, when the data do not contain the unambiguous answer sought by the investigator. (See Tukey's remarks above.) Note that the maneuver of pooling of sample and non-sample information on the parameter vector, or on the response, in order to salvage a weak study is a meta-analytic procedure - as described above. A brief exposition of modern regression methods, including Bayesian procedures, is provided below in sections 7-9 of this report.

e) The bias and variance of the estimates of the parameter vector $\underline{\beta}$ of the model and of linear and non-linear functions thereof, $f(\underline{\beta})$, such as the estimated response, say $\underline{x}^T\hat{\underline{\beta}}$, and the ratio, = $\theta = \beta_1/\beta_2$ $(=\alpha/\beta)$, are rarely (never?) adequately determined. (Indeed, the issue of the bias in the estimates of α/β has not hitherto been raised.) See Annex II, part 5 and Annex III, part 4. But, as Park and Snee (1983) observe, "The size of the confidence limits is inversely proportional to the quality of the data used to make the estimate and directly proportional to the amount of extrapolation involved. This important information is lost if the confidence limits and best estimates are not routinely reported. The width of the confidence interval is one of the best measures risk assessors, and risk managers, have to evaluate the quality of the estimates of potential risks. It is important to distinguish between those situations in which the risk is precisely estimated and those in which it is not."

f) As remarked above, there is an epistemological correlate to each of the ontological weaknesses discussed in section 6.1. We briefly review several of them. The epistemological correlate to the ontological failure of received practice to recognize the inherently multivariate nature of the deterministic part of the observed response is that the received dose-response models are not formulated, and the sample data are not represented, in terms of matrix theory. Matrix theory provides immediate and useful representations of the multivariate nature of the response and the mutual relations in the observations (rows) and variables (columns) in the sample information on dose-response. It is equally important that matrix theory also provides a parallel structure for the representation of the prior information on the level and precision of the observed response, and on the parameters of the model - either level or precision - thereby assuring that the sample and non-sample information on the matters at issue are isomorphic and hence can be readily compared and/or combined. (If the sample data are "weak" then it is necessary to "strengthen" them with non-sample information on the model parameters or functions thereof. See section 7.2.) For example, using the simple linear Normal theory model, the sample and a priori information on the parameter vector $\underline{\beta}$ of a model can be represented by the parallel, or isomorphic, forms

i) $\underline{y} = X\underline{\beta} + \underline{e}$ (sample) ii) $\underline{r} = R\underline{\beta} + \underline{v}$ (a priori)

The sample information on $\underline{\beta}$ is represented by the (n*k) observation matrix $[\underline{y}, X]$. The non-sample information on $\underline{\beta}$ by the (q*k) matrix of a priori constraint $[\underline{r}, R]$ where $q < k$. The (n*1) vector $\underline{e}$ represents the uncertainty in the observed response, $\underline{y}$. The (q*1) vector $\underline{v}$ represents the uncertainty in the non-sample information on $\underline{\beta}$ that is described by $\underline{r}$. (The equation i) will be discussed in section 7.1 and the equation ii) in section 7.2.) Therefore, the matrix notation facilitates model checking, discrimination, and validation, as well as model construction - including combining sample and non-sample information on the parameters and response.

g) There are also ontological implications of several of the epistemological maneuvers in received practice that many investigators do not seem to be sensitive to (several of these we have remarked before but they are important enough to repeat):

i) The linear model of a Binomial response, say $\eta = \alpha_0 + \alpha_1 D$ (see Fig. 13a) implies that the tolerance distribution of dose is rectangular (see Fig. 31b).

ii) The transformation $m_i \longrightarrow S_i$ for cell-survival data implies that the weight of the observation at $D = 0$ is infinite, say $w_1 = 10^6$.

iii) The maximum likelihood estimates of non-linear functions of the parameter vector $\underline{\beta}$ are inherently biased. (This is, of course, not true of linear functions of $\underline{\beta}$.)

iv) The least-squares estimates of α and β obtained from the Chapman cell inactivation plot for the LQ model of cell-survival data imply that the weights of the observations vary quadratically with dose, i.e., $w_i = D_i^{-2}$, $2 \leq i \leq n$ (see Fig. 2b). Items ii) and iv) describe the arbitrary adjustment of weights referred to in section 6.1.4 above.

6.2.3 Comparing rival models of a set of observations

It is important to repeat here that in every case in which the goodness-of-fit and consistency of a received model to given sets of data and to prior beliefs, respectively, were examined in a secondary analysis, we have been careful to also assess, respectively, both the goodness-of-fit and the consistency of the model of a rival hypothesis to the same sample data and a priori beliefs, using the same statistical concepts, methods and criteria. These concepts, etc., are 1) the (asymptotic) distributions of the respective sums of squared residuals, e.g., the Pearson chi-squared distribution, together with plots of the residuals and other regression diagnostics (vide infra) and 2) where possible, a measure of the degrees of invariance of the respective sample estimates of the parameter vectors of the two rival models between two different samples. This is the epistemological equivalent of the replication recommended by Zelen (1982) for clinical trials. Moreover, it helps to assure that the investigator is left with at least one model in each case. For, as Jeffreys (1961) has remarked, "Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule [or model] at all, whereas, the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts."

It is also the case, of course, that the "fitting" of a rival model to the sample data serves as a "control procedure" - to assure the reviewer that the sample data in the study selected for re-analysis does indeed contain determinate dose-response information and is not merely "garbage" - as in GIGO. In the latter case, the failure of the received model to "fit" the data - on statistically acceptable criteria - would carry no ontological implications. But it is also the case that, as Kuhn (1970b) has remarked, "In scientific practice the real confirmation questions always involve the comparison of two theories with each other and the world, not the comparison of single theories with the world." As remarked above, in comparing two rival models of a given sample of data on a given criterion of "fit", there are, of course, only three possible outcomes: 1) Neither model "fits"; 2) Both models "fit"; 3) One model fits, but the other does not. In every such comparison in the present study it was found that the models of currently received opinion, e.g., the LQ model, did not fit the data and prior beliefs as well as did the models of the respective rival hypotheses. For example, for the cell-survival data examined, the received linear-quadratic model (LQ) did not fit as well as the rival, Target theory, model, on measures of invariance of the estimates of model parameters between different sets of data, as well as on the statistically adequate measures of goodness-of-fit based on the various distributions of residuals (e.g., plots such as e_i vs $\hat{\mu}$) and their respective sums of squares (vide infra).

To pursue the issue of "fit" still further, it is most useful to distinguish two quite different ways in which the model and the data may be inconsistent: 1) numerical disagreement and 2) qualitative failure. Thus, although the data may not "reject" the model on statistically adequate criteria, that is, on the criteria of both aggregate and case statistics, the model may nonetheless be qualitatively - or topologically - incorrect. For example, the (local) sign of the curvature, k, of the dose-response curve of the model may be inconsistent with received opinions in a given region of response. Thus, for the two different sets of cell survival data that we used to compare the LQ with the rival Target model, the dose-response (survival) curve of the LQ model failed to exhibit "the shoulder" that is to be expected, a priori, on the basis of the generally accepted hypotheses of the possible mechanisms of biological responses to irradiation at the cellular level. (The absence of a "shoulder" implies the concomitant absence of either any redundant structures or of any "repair" of sublethal damage in the irradiated cells.) However, although the dose-response curves of the LQ models of these data were "concave-up", $k > 0$ in the region $0 \leq D \leq 2.0$ Gy, the cognate curves of the rival Target theory model (with the same number of parameters) were "concave-down", $k < 0$, in the same region; the Target model discerned the presence of a "shoulder" in the response data that the LQ model failed to perceive. Thus, the Target model was more consistent with the data, both quantitatively and qualitatively, than was the rival LQ model. In both cases we showed that the "shoulder" could be produced in the LQ survival curve as an artifact of the semi-log plot that is commonly used to display such curves. See part 7.10 and Annex II part 5 for further discussion and examples.

6.3 Model checking

Decisions on the adequacy of a model of a given set of data based on only qualitative appreciations, such as that a dose-response curve is "curvy" - or, "curvier" - (Fowler, 1984), or that an isoeffect plot is "straight" (Fowler, 1984) are, of course, not well-informed decisions although these are currently received practices. Instead, the quantitative checks on the adequacy of the "fit" of a model to a given set of data that were mentioned above are required. McCullagh and Nelder (1989) divide the methods of quantitative model checking into three categories: "1. tests of deviations in particular directions, 2. visual displays, 3. the detection of influential points. Category 1 methods are implemented by including extra parameters in the model, and testing whether their inclusion significantly improves the fit or not. Extra parameters might arise from including an extra covariate ... . By contrast, category 2 methods rely mainly on the human eye to detect pattern. Such methods take a successful model to be one that leaves a patternless set of residuals (or other derived quantities). The argument is that if we can detect pattern we can find a better model; the practical problem is that any finite set of residuals can be made to yield some kind of pattern if we look hard enough, so that we have go guard against over-interpretation." Nonetheless, category 2 methods are an important component in model checking.

Category 3 methods, when applied to Normal linear models, have become known as regression diagnostics. Their aim is the detection of data points that are in some sense particularly influential in the fitting; this may imply (but does not necessarily do so) that the inferences made from the model are especially heavily dependent on certain dependent response values or covariate values. It is important that the analyst should be aware of such points and their influence on his inferences. Influential points will often be outliers, in the sense of lying some distance from the main mass of data values; such outliers need to be identified so that they may be carefully checked in the original data, because they may, of course, be simply wrong values, the result of transcription or recording error."

7. Statistical methods II

"It is important to realize that the statistical properties of a model remain the same, irrespective of the discipline or application of the model."

D. Ratkowsky, 1990

"... it is a function of statistical method to emphasize that precise conclusions cannot be drawn from inadequate data."

E. Pearson and H. Hartley, 1970

"A regression is constructed using prior knowledge, data, and a fitting (estimation) process of some form. It is important to know when the resultant regression depends heavily on a small part of the prior knowledge, on a small part of the data or on the exact choice of model or fitting process."

R. Welsch, 1986

"The fact that a small subset of the data can have a disproportionate influence on the estimated parameters or predictions is of concern to users of regression analysis, for, if this is the case, it is quite possible that the model estimates are based primarily on this data subset rather than on the majority of the data."

Belsley, Kuh, and Welsch, 1980

"An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations. One obvious means for examining such an impact is to delete each row, one at a time, and note the resultant effect on the various calculated values. Rows whose deletion produces relatively large changes in the calculated values are deemed influential."

Belsley, Kuh, and Welsch, 1980

The preceding discussions have served, in part, to motivate the reader. They have made it apparent that the issues of ontological status and epistemological practice in radiobiological modelling that have been raised so far in this report will be much clarified by interpolation of another and lengthier digression on statistical concepts, methods, and criteria. Such a digression will also enable us to illustrate the further discussion of the issues much more concisely. For a fuller exposition of the matters discussed below the interested reader may consult the following textbooks: Draper and Smith, 1981; Montgomery and Peck, 1982; Gilchrist, 1984; Cook and Weisberg, 1982; McCullagh and Nelder, 1989; Leamer, 1978; Theil, 1971; Finney, 1971a; Hosmer and Lemeshow, 1989; Weissberg, 1985; Belsley, Kuh and Welsch, 1980.

7.1 Estimation and inference. Model-sample interactions. Methods and criteria. Parametric regression. Regression diagnostics. Robust procedures. Experimental design criteria. Non-parametric regression.

7.1.1 Parametric regression.

It will be convenient to illustrate the fruitfulness of matrix methods in the conceptualization of dose-response models by taking the more familiar example in which the response, y, has a (conditional) Normal distribution, the so-called Normal-theory model, rather than a (conditional) non-Normal distribution, e.g., Binomial or Poisson. The extension to cases in which y has a non-Normal distribution is, of course, quite straight-forward and is well described in Dobson (1982) and McCullagh and Nelder (1989). However, it is important to remark here an important operational distinction between models of data in which the response has a Normal distribution and those in which that distribution is non-Normal. In the former case, the ordinary least squares (OLS) estimates of the parameter vector, say $\underline{\beta}$, are uniquely determined from the sample by a single matrix inversion. But if the response, $\underline{y}$, has a (conditional) non-Normal distribution then the sample estimates of the parameter vector $\underline{\beta}$ depend upon a construct, say $\underline{y}^{\#}$, the pseudo-response vector, often called the <u>adjusted dependent variable</u>, as well as the weight matrix, W. Both $\underline{y}^{\#}$ and W are functions of the estimated response $\hat{\underline{y}}$, and hence of the parameter estimate $\hat{\underline{\beta}}$. Thus, the estimation of $\underline{\beta}$ is an iterative procedure, that is, the estimates of the parameter vector are determined by a sequence of matrix inversions - the iteratively reweighted least squares estimates (IRLS). Thus, the latter estimates are not determined uniquely by the data but also, to varying degrees, by arbitrary a priori choices of a) the initial estimates - the starting values for the iterations - of the parameters, b) the stopping rule - and hence the number of iterations - and, c) the approximations used in the algorithm. The sensitivity of the parameter estimates to each of these a priori choices should be determined for each model-data combination - as recommended by Belsley et al, 1980.

Using the Normal theory model as an example we now consider the two problems of a) representing the sample information on the multivariate dependence of the response in terms of a regression model with a parameter vector $\underline{\beta}$ and b) combining sample with non-sample information on $\underline{\beta}$, which can be represented by a model of similar form. In the matrix calculus a linear regression model of the observation matrix, $[\underline{y}, X]$ on a sample of size n is represented as $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$ where $\underline{y}$ is an (n*1) response matrix (vector), $\underline{\varepsilon}$ is an (n*1) noise matrix of zero mean and constant variance, σ^2; $\underline{\varepsilon}$ is Normally distributed: $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n)$, where I_n is the n*n identity matrix. X is an (n*k) matrix of treatment variables and $\underline{\beta}$ is a (k*1) vector of unknown parameters; $[\underline{y}, X]$ is an n*(k+1) matrix of observations. In this exposition we assume that the model includes an intercept and hence the matrix X includes a unit column vector and p column vectors of predictor variables, X_j, $1 \le j \le p$. We further stipulate that the sample size, n, is adequate: "In order to have adequate degrees of freedom for error, giving reasonable power for significance tests and a meaningful residual analysis, the size of the estimation set, n should be greater than k + 10 (or k + 15) where k is the largest number of coefficients one believes will be required to describe the response." (Snee, 1977).

The ith row of the above equation may be written as

$$y_i = \underline{x}_i^T\underline{\beta} + \varepsilon_i.\ 1 \le i \le n$$

The full matrix equation for the model is

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{y}$ and X are known and observable and $\underline{\beta}$ and $\underline{\varepsilon}$ are unknown and unobservable. The fundamental features of the model can be described in a fourfold table:

	known and observable	unknown and unobservable
stochastic	y	$\underline{\varepsilon}$
deterministic	X	$\underline{\beta}$

The least squares estimator, $\hat{\underline{\beta}}$, of the unknown parameter vector $\underline{\beta}$ is obtained by a single matrix inversion from the so-called <u>normal equations</u>, $X^TX\bar{\underline{\beta}} = X^T\underline{y}$, which are obtained as solutions of $\partial[(\underline{y} - X\bar{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})]/\partial\bar{\underline{\beta}} = \partial[\Sigma e_i^2]/\partial\bar{\underline{\beta}} = \underline{0}$:

$$\bar{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$$

$\hat{\underline{\beta}}$ is an unbiased, minimum variance, maximum likelihood estimate of $\underline{\beta}$ with covariance matrix

$$\mathrm{Var}(\hat{\underline{\beta}}) = \sigma^2(X^TX)^{-1}$$

where RSS, the so-called residual sum of squares is defined as

$$RSS = (\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}}) = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \underline{x}_i^T\hat{\underline{\beta}})^2 = e^Te = \Sigma e_i^2$$

$y_i = \underline{x}^T\bar{\underline{\beta}}$ is the response predicted by the model at $\underline{x}_i^T$ and $e_i = (y_i - \bar{y})$ is the ith <u>residual</u>. If the model "fits" (vide infra) then $\hat{\sigma}^2 = RSS/(n-k)$ is an unbiased estimate of σ^2. (<u>N.B.</u>: Since the prediction, $y_i = \underline{x}_i^T\bar{\underline{\beta}}$, is in terms of the <u>unknown</u> parameter vector, $\underline{\beta}$, of which $\hat{\underline{\beta}}$ is the estimate obtained from a finite sample, the regression model may be said to "explain" the <u>known</u> set of observations, (y_i, x_i^T), $1 \le i \le n$, in terms of the <u>unknown</u> parameter vector, $\underline{\beta}$, of the regression model. Or, "It can be said without paradox that scientific explanation is, ..., the reduction of the known to unknown." K. Popper (19665a/1965b).)

The model can be expressed in terms of the sample estimates of $\underline{\beta}$ and $\underline{e}$: $\underline{y} = X\hat{\underline{\beta}} + \underline{e}$. The vector of estimated responses, $\hat{\underline{y}}$, and the vector of residuals $\underline{e}$ can be expressed in terms of an (n*n) projection matrix, the so-called hat matrix, H:

$$\hat{\underline{y}} = X\hat{\underline{\beta}} = X(X^TX)^{-1}X^T\underline{y} = Hy$$

and

$$\underline{e} = \underline{y} - \hat{\underline{y}} = [I_n - H]\underline{y}$$

where I_n is the (n*n) identify matrix.

The mean and variance of $\underline{e}$ are given by

$$E(\underline{e}) = \underline{0} \text{ and } Var(\underline{e}) = \sigma^2(I-H).$$

At the i^{th} level of treatment, $\underline{x}_i^T$, $1 \leq i \leq n$, we have $\partial\hat{y}_i/\partial y_i = h_i$, where $h_{ii} \equiv h_i = \underline{x}_i^T(X^TX)^{-1}\underline{x}_i$ is the i^{th} diagonal element of the hat matrix H. Also, we find that $E(\hat{y}_i) = y_i$, $E(e_i) = 0$. And, $Var(\hat{y}_i) = Var(\underline{x}_i^T\hat{\underline{\beta}}) = \underline{x}_i^TVar(\hat{\underline{\beta}})\underline{x}_i = \underline{x}_i^T\sigma^2(X^TX)^{-1}\underline{x}_i = h_i\sigma^2$ and $Var(e_i) = \sigma^2(1-h_i)$. (N.B. The e_i are correlated but the ϵ_i are not.) Note that $Var(\hat{y}_i)$ depends quadratically on $\underline{x}_i^T$. In general, whatever the form of the distribution of ϵ_i, it will be found that the variance of the predicted response is a minimum at the centroid of the distribution of the sample and a maximum at the extremes of that distribution. Hence, interpolation is more precise, as well as more accurate, than extrapolation.

The residuals, e_i, and the hat matrix diagonals, h_i, are case statistics known as regression diagnostics. These are a set of n measures (one for each case, or observation, in the sample) that will "... readily identify observations that are not well-explained by the model, as well as those dominating some important aspect of the fit" (Pregibon, 1982). "Large" values of the standardized residuals, $e_i^* = e_i/\sqrt{Var(e_i)} = e_i/\sigma\sqrt{(1-h_i)}$, identify those observations, "not well explained by the model". These are usually referred to as "outliers". "Large" values of h_i identify those observations that may be, "dominating some important aspect of the fit". These are usually referred to as "high leverage" observations. In Normal theory models, h_i is a measure of the "distance" of the i^{th} observation from the centroid of the data. The criteria for identifying observations as "outlying" and/or "high leverage" are commonly taken to be $e_i^* > 3.0$ and $h_i > 2k/n$. The "cut-off" for e_i^* is based on the fact that if the model "fits" e_i^* is distributed approximately as a unit Normal deviate: $e_i^* \sim N(0,1)$. In practice, the identification of outlying observations is based on additional information such as can be provided by probability plots of the residuals. And often a more conservative criterion is employed, say $e_i^* > 2.0$. The cut-off of 2k/n for h_i is selected because $\Sigma h_i = k$ and hence the average value of h_i is $\bar{h} = k/n$. It is important to note that if $h_i = 1.0$, then the i^{th} observation completely determines the sample estimate of one of the parameters of the model.

RSS is an aggregate statistic that provides an overall measure of the goodness-of-fit of the model to the original sample. However, it provides a biased (optimistic) estimate of the predictive performance of the model in new data. A reduced-bias estimate of the predictive performance of the model in new data is provided by the PRESS statistic (the predictive sum of squares) to be discussed briefly below and at greater length in section 9.3 In exploiting the RSS as a measure of the fit, or concordance, of the model to a set of data, it is necessary to assume that the sampling distribution of the RSS - or a given function of the RSS, say f(RSS) - is well-approximated by that of a known random variate, such as the F-distribution for models of data in which the response has a Normal distribution, or the Pearson chi-squared distribution for models of data in which the response has a Binomial or Poisson distribution. In these latter cases the appropriate residuals (to be squared and summed) are the Pearson chi-squared, χ_i, or deviance, d_i, residuals. The respective residual sums of squares are $RSS = \Sigma\chi_i^2 = \chi^2$ and $RSS = \Sigma d_i^2 = D$, the deviance. (N.B. For the Normal theory models, the deviance and chi-squared residuals are just the e_i described above.) For the models of Binomial and Poisson responses, respectively, they take the following forms (See Table 3a for additional forms of residuals).

Binomial $\chi_i = (r_i - n_i\hat{\pi}_i)/\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$

where r_i is the number of responders out of n_i at risk at $\underline{x}_i^T$ and $\hat{\pi}_i$ is the cognate proportion estimated from the model with parameter estimate $\hat{\underline{\beta}}$.

Poisson. $d_i = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{[2\{y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i\}]}$.

$\text{sgn}(y_i - \bar{\mu}_i)$ is just the signum function; it is +1 or -1 according to the sign of $(y_i - \bar{\mu}_i)$. For k-parameter generalized linear models of samples of size n, the sampling distributions of χ^2 and D are approximated asymptotically by the Pearson chi-squared distribution on (n-k) degrees of freedom.

As well as the "fit" of a given model to the data, it is necessary to assess the stability of the estimates of the parameters, $\underline{\beta}$, and of functions of the parameters such as the response $\underline{x}_i{}^T\underline{\beta}$ (a linear function) and ratios such as $\theta = \beta_1/\beta_2$ (a non-linear function). In general, the stability of an estimate is assessed by comparing it with its standard deviation on the assumption that the sampling distribution of the ratio is sufficiently well-approximated by either the Student's t or the Normal distribution. For example, for a k-parameter model of a sample of size n, $t_j = \hat{\beta}_j/\sqrt{\text{Var}(\hat{\beta}_j)}$ is distributed approximately as Student's t on (n-k) degrees of freedom. Typically, the variable X_j is deleted from the model if $|t_j|$ does not exceed the 0.975 quantile of the Student's t distribution.

But, "In general, the model is not likely to be a satisfactory predictor unless the range of fitted values y_i [say $\hat{y}_{max} - \bar{y}_{min}$] is large relative to their average estimated standard error $\sqrt{k\sigma^2/n}$ where σ^2 is a model-independent estimate of the error variance" (Montgomery and Peck, 1982). As a rule of thumb, for a "useful model", each parameter estimate must exceed its estimated standard error by a factor of at least 8: $\hat{\beta}_j/\sqrt{\text{Var}(\bar{\beta}_j)} \geq 8$, $0 \leq j \leq p$. This is equivalent to the so-called Box-Wetz criterion (Box and Wetz, 1973). It is also consistent with the g-value criterion for "useful" simple probit models, $z = \beta_0 + \beta_1 x$, where z is the probit transform $z = \Phi^{-1}(\pi)$, $0 \leq \pi \leq 1$ and which are identified by $g = 1.96^2 \text{Var}(\hat{\beta}_1)/\bar{\beta}_1{}^2 \leq 0.05$ (Finney, 1971b).

It is important to note that if $\hat{\beta}_j/\sqrt{\text{Var}(\hat{\beta}_j)} < 1$ then the variable Xj should always be deleted from the model since the increase in bias (of the estimate of $\underline{\beta}$) due to its omission is less than the increase in variance due to its inclusion, hence the mean squared error is minimized (Montgomery and Peck, 1982).

Least squares is not a robust procedure in that modest changes in the data can produce substantial changes in the estimates and inferences obtained by this procedure. However, least squares combined with the procedures of regression diagnostics to be discussed below is a robust procedure, since cases that are not well-explained by the model, or that will dominate some important aspects of the fit, will be identified, and their effects on the estimates and inferences that are to be made from the model can be assessed and, often, reduced.

As remarked above, overall goodness of fit of model and data is described by functions of the squared residuals. For Normal theory model the residual is $\hat{e}_i = (y_i - \hat{y}_i)$. The sum of squared residuals is $\Sigma e_i{}^2 = \Sigma(y_i - \hat{y}_i)^2$ and one function of $\Sigma\hat{e}_i{}^2$ that is of interest is the square of the multiple correlation coefficient, R:

$$R^2 = 1-(\Sigma(y_i - \bar{y}_i)^2)/(\Sigma(y_i - \overline{y})^2)$$

where $\overline{y} = \Sigma y_i/n$ is the mean response and R is just the correlation coefficient $1.0 \leq R \leq 1.0$ of observed, y_i, and estimated, $\hat{y}_i$, levels of the dependent variable. R^2 describes the proportion of the observed variation in y_i about the mean $\overline{y}$, in the sample data, that is accounted for by the linear model, $y = X\hat{\underline{\beta}} + \underline{e}$.

The adjusted multiple correlation coefficient squared, $R_a{}^2$, describes the correlation of y_i and $\bar{y}_i$ in new data.

$$R_a{}^2 = 1 - [(n-1)/(n-k)](1-R^2)$$

Note that $R_a{}^2 < R^2$; like the PRESS statistic that is to be described below, $R_a{}^2$ provides a measure of the degradation of the predictive performance of the model, $y = X\hat{\underline{\beta}}$, in new data. Thus, a "useful" Normal theory model can be (partly) identified by the value of $R_a{}^2$; for such models $R_a{}^2 \geq 0.80$.

The correlation coefficient, R, describing the correlation of observed, y_i, and expected, $\hat{y}_i$ (or $\hat{\mu}_i$), responses, is equally useful describing the concordance of models of non-Normal data. A graphical description of the correlation between y_i and $\bar{y}_i$ is given by the so-called fit-observation plot of y_i vs $\bar{y}_i$ (Gilchrist, 1984). N.B.: It is seen above that for a linear model R^2 and $\Sigma e_i{}^2$ are directly related so that the model for which Σe_i^2 is a minimum is also the model for which R^2 is a maximum. However, for "... non-linear models the relation between $[R^2]$ and $\Sigma \hat{e}_i^2$ ceases to hold, so that they become different criteria ... Thus, the model ... [with the largest R^2] is not necessarily the best fitting model ... The obvious ways to avoid being misled are to always plot the fit-observation diagram and to evaluate $\Sigma \hat{e}_i^2$ alongside $[R^2]$." (Gilchrist, 1984). Examples are provided in Figs. 18a for a Binomial distribution of response and 24a for a Poisson distribution of response. As in the case of the probability plot (a graphical assessment of the validity of the selected model form for the <u>distribution</u> of the <u>random</u> part, $\underline{\varepsilon}$, of the observed response) if the plot of y_i vs $\hat{y}_i$ is well-described by a straight line, this suggests that the selected model form for the <u>deterministic part</u>, $\underline{\mu}$, of observed response is <u>not invalid</u>. It is also useful in discriminating between non-nested models, for which discrimination procedures based on the decrements in sums of squared residuals are not valid. Additional procedures for discrimination between nested rival models and between non-nested rival models are presented in sections 7.7 and 7.8.

The log-likelihood function, $\ln L(\hat{\underline{\beta}})$, evaluated at $\hat{\underline{\beta}}$, is an appropriate overall measure of goodness-of-fit when $\hat{\underline{\beta}}$ is a <u>maximum-likelihood</u> estimate of $\underline{\beta}$. $\ln L(\hat{\underline{\beta}})$ is useful in assessing the significance of sub-sets of predictor variables in a procedure known as the likelihood ratio test. In particular, Wilks (1962) shows that $\lambda = -2\ln[L(\hat{\underline{\beta}})/L(\beta_0)]$ is distributed asymptotically as $\chi^2(p)$ where $p = k-1$, the number of explanatory variables in the model. The likelihood ratio index, $\rho^2 = 1-\ln L(\hat{\underline{\beta}})(\beta_0)$, sometimes called McFadden's ρ^2, can be used as a <u>pseudo-R^2</u> to measure the goodness-of-fit of a model. But it must be remarked that although, $0 \leq \rho \leq 1$, values $0.2 \leq \rho^2 \leq 0.4$ describe <u>extremely good fits</u>, whereas only values $0.9 \leq R^2 \leq 1.0$ describe good fits. (Hensher and Johnson, 1981).

Box and Hunter (1965) have remarked that, "All the information relating to the possible <u>inadequacy</u> of a tentatively entertained model is contained in the residuals. Plots of the residuals, therefore, can reveal particular aspects of the model that should be improved. Consequently, as a matter of course, residuals should always be plotted in any way that might shed light on pertinent questions." We note also that, "The idea of inspecting residuals is very old, but the systematic calculation of residuals, particularly from extensive data, has become practicable only recently: their thorough graphical analysis is feasible only with a suitable computer graphical output device." (D. Cox and E. Snell, 1968). For a k-parameter model of a sample of size n Snee (1977) recommends that the sample size include at least $n = k+10$ observations for "... a meaningful residual analysis." (However, we present in this report several examples of "meaningful" residuals plots for k-parameter models of samples for which $n < k + 10$.)

Figures 5-8 present an analysis of the residuals for three rival versions of the LQ model of the mutagenesis data of Fig. 1. The data are Tradescantia from Sparrow et al, 1972; the data in Fig. 1 provided the estimate of α/β in Table <u>5.1</u> of NCRP 64, 1980.) Two of the LQ models are versions of the standard power-Normal model of Snee (1986) and Snee and Irr (1981), $m_i^\lambda = (\beta_0 + \beta_1 D_i + \beta_2 D_i^2)^\lambda + \epsilon_i$ for $\lambda = 0$ (Fig. 5) and $\lambda = 0.5$ (Fig. 6). (Power-Normal models are frequently used for bio-assays when the data are counts and the experiment is well understood. See also Carroll and Ruppert, 1984, 1988.) The third is the standard Poisson linear model, $m_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$ (Figs. 7 and 8). Figures 5 and 6 are plots of standardized residuals, e_i^*. Figures 7 and 8 are the Pearson chi-residuals, χ_i, for the Poisson distribution. The residuals for the power-Normal model and for the Poisson linear model have the same general form: $e_i^* = (y_i - \hat{\mu}_i)/\sqrt{Var(e_i)}$. For the power-Normal model, $Var(\varepsilon_i) = \sigma^2$; for the Poisson model, $Var(\varepsilon_i) = \mu_i$.

Note that the index numbers in the abscissa of the plots of Figs. 5b, 6b, 8a, 8b are simply the (ordinal) row numbers of the observations in the matrix $[\underline{y}, X]$. For these data each set of 5 consecutive numbers refers to a <u>different</u> experiment in the set of four that were pooled to form the Sparrow et al (1972) data of Fig. 1.

Figure 6c describes the results of the empirical method of Snee (1986) for identifying the <u>form of the distribution</u> of the random part of the response for these data, either Poisson or Negative Binomial: For the correct choice of λ the residuals of the model will have a Normal distribution - as in Fig. 6a. It is evident from Fig. 6c that the errors in the Sparrow et al data have a Poisson distribution ($\lambda = 0.5$) rather than Negative Binomial ($\lambda = 0$) as is assumed in NCRP 64 (1980). Note that the estimate of λ is obtained by an iterative procedure and that, for a given value of λ, the sample estimate of the parameter vector $\underline{\beta}$ is also obtained by another iterative method. The power-Normal models are used routinely to analyze similar data on chemical mutagens. It is important to note that, in sharp contrast to the Sparrow et al data of Fig. 1, the data in the well-designed experiments on mutagenesis include equal numbers of replicates (say 4) at each level of the mutagen. For such data, the sample estimates of the parameter vector $\underline{\beta}$ is fairly insensitive to changes in λ, only the covariance matrix changes appreciably with λ

The method of identifying the form of the distribution of the random part of the model of the Tradescantia data that is described in 6c is complementary to - and in practice, prior to - the probability plots of Figs. 5a and 6a. (N.B.: The Fig. 6c nicely illustrates the general principles of the method of embedding for discrimination between two non-nested rival models. The method can be used, as in Fig. 6c, to discriminate between rival models of the random part of a model; it can also be used to discriminate between rival models of the deterministic part - as described for instance in Muirhead and Darby, 1987 (relative vs absolute risk models of radiation-induced cancer in human populations). "The basic idea of embedding is to devise a new model M2 that has both the non-nested models M_0 and M_1 as special cases." ... "The problem of identification is thus turned into one of estimation" W. Gilchrist (1984). See also sections 7.7, 7.8, and 14.2.)

The F_e-plot of Fig. 2a described a common example of a set of experimental isoeffect data in which the sample estimate of the parameter $\underline{\beta}$ of the LQ regression model is dominated by the single observation at d = D/N = 15 Gy. The residuals plots of Figs. 5b, 6b, and 8a describe an example in which the sample estimate of the parameter vector of the LQ dose-response model of the pooled observations is dominated by a single set of five observations - index #'s 6-10 -that comprise experiment #5 in the pooled data. The observations in experiment #5 have greater weight than do the observations in the remaining three experiments and hence will dominate the parameter estimates of models of the pooled data. The Sparrow et al study provides example of an analogue of Simpson's paradox: "... when true effects vary from study to study the size and even the direction of the combined result may depend heavily on such extraneous features as which studies were largest and may hence tend to dominate the analysis." (Halvorsen, 1986). A complete analysis of these data is presented in Annex III, part 5.

As remarked above in section 6.1.4, it is, of course, the case that not all observations are equal. Some observations include more information than others; these observations carry greater weight. Therefore, let us now examine still further the weight, w_i, of an observation, $(y_i, \underline{x}_i^T)$. It will greatly simplify the exposition to restrict it to linear models, that is, to models linear in the parameter vector, $\underline{\beta}$, the so-called generalized linear models described in section 4.

Let $\underline{x}_i^T\underline{\beta} = \eta_i$, the (so-called) linear-predictor. The (so-called) expected value of the response at $\underline{x}_i^T$ is $E(y_i) = \mu_i$. For the simple linear Normal theory model we have $y_i = \mu_i + \epsilon_i$ and $\mu_i = \eta_i$. The inherent weight of the ith observation is

$$w_i = \frac{1}{Var(y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

For normal theory model at issue it is obvious that $\partial\mu_i/\partial\eta_i = 1$ and $Var(y_i) = \sigma^2/n_i$ where n_i is the number of observations at $\underline{x}_i^T$. Thus, $w_i = n_i/\sigma^2$. An estimate is $\hat{w}_i = n_i/\hat{\sigma}^2$. The corresponding aggregate measure of fit is $RSS = \Sigma w_i e_i^2$.

As an example of a non-Normal distribution consider the Poisson distribution in which $Var(y_i) = E(y_i) = \mu_i/n_i$, we have $w_i = n_i/\mu_i$. Similarly, for a probit model of a response with a Binomial distribution in which $Var(y_i) = \pi_i(1-\pi_i)/n_i$ $0 \leq \pi_i \leq 1.0$ where π_i is the conditional probability of response at $\underline{x}_i^T$, we have $w_i = n_i f(z_i)^2/\pi_i(1-\pi_i)$ where $f(z_i)$ is the ordinate of the Normal density function at z_i: $f_i = (1/\sqrt{2\pi})\exp(-z_i^2/2)$.

The generalized linear models of the Poisson responses of interest in this report have the forms $m_i = \underline{x}_i^T\underline{\beta}$ for mutagenesis and carcinogenesis, and $m_i = \exp(\underline{x}_i^T\underline{\beta})$ for cell-survival where m_i is the Poisson rate parameter and $1 \leq i \leq n$, the sample size. These are the linear and log-linear Poisson models, respectively.

The generalized linear models of the Binomial responses of interest in this report have the forms, $z_i = \underline{x}_i^T\underline{\beta}$ where $z_i = \Phi^{-1}(\pi_i)$ for the <u>probit</u> model and $z_i = \log[\pi_i/(1-\pi_i)]$ for the <u>logit</u> model, $0 \leq \pi_i \leq 1$, $1 \leq i \leq n$, and $\Phi(.)$ is the Normal distribution function.

The a priori information on the weights, w_i, $1 \leq i \leq n$, of the observations $(y_i, \underline{x}_i^T)$ can be usefully included by transforming both sides of equation (Theil, 1971; Johnston, 1972):

$$P\underline{y} = PX\hat{\underline{\beta}} + P\underline{\epsilon}$$

where $P^TP = V^{-1} = W$ where W is an (n*n) diagonal matrix with diagonal elements, w_i, $1 \leq i \leq n$. Then

$$\hat{\underline{\beta}} = (X^TWX)^{-1}X^TW\underline{y}$$

and

$$Var(\hat{\underline{\beta}}) = \sigma^2(X^TWX)^{-1}$$

Note that the hat matrix is $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ and the diagonal elements, h_i, now represent both the "distance" of the i^th observation from the centroid of the data <u>and</u> its inherent weight. Here $W^{1/2}$ denotes the diagonal matrix for which the diagonal elements are the square roots, $\sqrt{w_i}$ of the corresponding elements of the matrix W.

For distributions of the random part of the response in which the variance is a function of the expected value such as the Binomial and Poisson, the estimates of the matrices, W and $\underline{\beta}$ are obtained interatively. In this procedure the dependent variable is <u>not</u> y but $y_i^{\#}$, the so-called <u>adjusted</u> dependent variable, where $y_i^{\#} = \eta_i + (y_i - \hat{\mu}_i)(d\eta/d\mu)_i$. The procedure is iterative because both $y_i^{\#}$ and W depend on the fitted values, $\bar{\mu}_i$, for which only current estimates are available. See McCullagh and Nelder, 1989.

For the probit model the weights have the form (Finney, 1971b):

$$w_i = n_if_i^2/\hat{\pi}_i(1-\hat{\pi}_i)$$

where $f_i = (1/\sqrt{2\pi})\exp(-z_i^2/2)$ and n_i is the number of risk at $\underline{x}_i^T$. For the logit model we have

$$w_i = n_i\hat{\pi}_i(1-\hat{\pi}_i)$$

The P-transformation of the observation matrix $[\underline{y}, X]$ described above ($\underline{y}$ $P\underline{y}$, X   PX, $\underline{\epsilon}$ $P\underline{\epsilon}$) is often useful in the subsequent computation of the regression diagnostics that are required for the model validation and discrimination procedures to be discussed below. The weight matrix $W = P^TP$ is that obtained at the final iteration. (See also Annexes II-IV.)

For $W = I_n$, the (n*n) identity matrix, the estimate of $\underline{\beta}$ that is obtained when the sample is perturbed by deletion of the i^th observation $(y_i, \underline{x}_i^T)$ giving an estimation sample of size (n-1) is the <u>row-deleted estimate</u>, $\hat{\underline{\beta}}_{(i)}$, where

$$\hat{\underline{\beta}}_{(i)} = [X_{(i)}^TX_{(i)}]^{-1}X_{(i)}^T\underline{y}_{(i)},\ 1 \leq i \leq n$$

and

$$Var(\hat{\underline{\beta}}_{(i)} = \sigma_{(i)}^2(X_{(i)}^TX_{(i)})^{-1}.$$

Alternatively, we may estimate $\hat{\beta}_{(i)}$ from the key diagnostics, e_i and h_i:

$$\hat{\underline{\beta}}_{(i)} = \hat{\underline{\beta}} - (X^TX)^{-1}\underline{x}_ie_i/(1-h_i)$$

where h_i is the i^th diagonal element of the <u>hat matrix</u>, H, and e_i is the i^th residual.

An alternative procedure for obtaining the row-deleted estimates, $\hat{\underline{\beta}}_{(i)}$, for generalized linear models is based on the <u>mean shift outlier model</u>. (Belsley et al, 1980; Cook and Weisberg, 1982) It is most convenient to illustrate this method for the Normal model which we may write as

$$y_j = \underline{x}_j^T\hat{\underline{\beta}} + e_j,\ 1 \leq j \leq n.$$

One way to assess the possibility that the i^th case is an <u>outlier</u> is to augment the deterministic part of the above model to give

$$y_i = \underline{x}_j^T\hat{\underline{\beta}} + \phi d_j + e_j,\ 1 \leq j \leq n$$

where $d_j = 1$, if $i = j$ and $d_j = 0$ if $i \neq j$. It can be readily verified that the maximum likelihood (IRLS) estimator of $\underline{\beta}$ for the augmented model is equal to $\hat{\underline{\beta}}_{(i)}$, the row-deleted estimate of $\underline{\beta}$ for the initial model. The estimate of $\underline{\beta}_{(i)}$ obtained by this procedure does not have the weaknesses of the cognate estimates calculated at final iteration from the weighted observations, $[P\underline{y}, PX]$, described above. Note also that the ratio, $\hat{\phi}/\sqrt{Var(\hat{\phi})}$, gives the j^th Studentized residual, (RSTUDENT) $e_i^{\#} = e_i/\hat{\sigma}_{(i)}\sqrt{(1-h_i)}$, where $\sigma_{(i)}$ is the row-deleted estimate of σ, which is distributed as a unit Normal deviate, that is, $\phi/\sqrt{Var(\phi)}$ – N(0,1).

The diagnostic $\underline{\beta} - \underline{\beta}_{(i)} = (S^{T}X)^{-1}\underline{x}_i e_i/(1-h_i)$ is termed DFBETA (see Belsley et al, 1980). $|\,\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}}\,|$ may be "large" if either e_i or h_i - or both - are large. Observations, y_i, $\underline{x}_i^{T}$, for which $\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}}$ is large are termed influential. Models for which $\bar{\underline{\beta}}_{(i)} - \bar{\underline{\beta}} \simeq \underline{0}$, i.e., for which the parameter estimate is invariant under deletion are useful (N.B.: "... in an adequate model, constants stay constant when variables are varied" Box, Hunter, and Hunter, 1978.) A standardized measure of influence is Cook's distance, D_i. For the k-parameter Normal theory model,

$$D_i = (\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}}^{T})(X^{T}X)(\bar{\underline{\beta}}_{(i)} - \hat{\underline{\beta}})/k\sigma^2$$

or

$$D_i = (\bar{\underline{y}}_{(i)} - \hat{y})^{T}(\hat{\underline{y}}_{(i)} - \hat{\underline{y}})/k\sigma^2.$$

Typically, the case (or cases) for which D_i is "large" are of interest. For the Normal theory model $D_i = 1$ identified observations for which deletion will move the estimate $\hat{\underline{\beta}}_{(i)}$ to the – 0.50 ellipsoid on $\underline{\beta}$. These are identified as "influential". Cook's D_i may be written in terms of the diagnostics e_i and h_i:

$$D_i = e_i^{\#2}h_i/k(1-h_i).$$

The standardized components of DFBETA (Belsley et al, 1980) are

$$\text{DFBETAS}_{ji} = (\bar{\beta}_j - \bar{\beta}_{j(i)})/\sigma_{(i)}\sqrt{}(X^{T}X)_{jj}^{-1}.$$

"Large" values of DFBETAS_j are identified as those for which $\text{DFBETAS}_{ji} > 2/\sqrt{}n$.

Although we have discussed the measures D_i and DFBETAS_{ji} only for Normal theory models, they may be readily generalized in obvious ways to models of data in which the response has Binomial and Poisson distributions. See Annex II parts 3 and 5, for example.

The largest, h(n), of these of hat matrix diagonals for a given model of a given model of a given set of data of size n provides a useful discriminator in identifying estimated responses, say $\underline{x}^{*T}\bar{\underline{\beta}}$ at $\underline{x}^{*T}$, that are interpolations, and those that are extrapolations of the model. For $\underline{x}^{*T}(X^{T}X)^{-1}\underline{x}^{*} > h_{(n)}$, $\underline{x}^{*T}\bar{\underline{\beta}}$ is an extrapolation; for $\underline{x}^{*T}(X^{T}X)^{-1}\underline{x}^{*} \leq h_{(n)}$, $\underline{x}^{*T}\hat{\underline{\beta}}$ is an interpolation. (Montgomery and Peck, 1982) N.B.: It must be noted that in extrapolation, the unbiased least-squares estimator of the parameter $\underline{\beta}$ of a model may not yield the best predictions of response in a minimum mean-squared-error sense. Biased estimators are usually better (See Allen and Jordan, 1982). These are sometimes referred to as non-least-squares estimators; an example of such an estimator is the Ridge regression estimator discussed below in 7.2.1.

The estimate of the response predicted at $\underline{x}_i^{T}$ when the ith observation is deleted from the estimation sample is $\hat{y}_{(i)} = \underline{x}_i^{T}\bar{\underline{\beta}}_{(i)} = (\hat{y}_i - h_i\hat{y}_i)/(1-h_i)$.The squared correlation of $\hat{y}_{(i)}$ and y_i, a measure of the predictive performance of the model in new data is termed the cross-validation R^2, or R_{cv}^{2}. Note that $R_{cv}^{2} < R^2$.

The ith predicted residual is $e_{(i)} = y_i - \bar{y}_{(i)} = e_i/(1-h_i)$. The so-called jackknife validation statistic is $\text{PRESS} = \Sigma e_{(i)}^{2} = \Sigma e_i^{2}/(1-h_i)^2$. PRESS is also a useful aggregate measure of the goodness-of-fit of the model to new data and hence of the degree of which the model may be generalized. Note that the set of residuals, e_i, $1 \leq i \leq n$, provides one measure, RSS, of how well the model "fits" the sample in hand. Also they may be used to assign weights to observations, to suggest data transformations, or to suggest including additional covariates in the linear predictor, η_i. When weighted by simple function, $(1-h_i)^{-1}$, of the cognate hat matrix diagonal, the set of e_i provides a measure, PRESS, of how readily the model can transcend the sample. Always, we have PRESS > RSS, but for an adequate model, the difference is small; PRESS = RSS.

Comparison of PRESS and RSS, e.g., PRESS/RSS, or PRESS-RSS, provides a measure of the degree of the degradation of the predictive performance of the model in new data. ("Users have often been disappointed by ... multiple regression equations that 'forecast' quite well for the data on which they were built. When tried on fresh data, the predictive power of these [equations] fell dismally." Mosteller and Tukey, 1977. That is, the form and/or parameters of the model were not invariant between the two sets of data; the consequences are sometimes described as, "the regression of the regression".) The latter measure, PRESS-RSS, is a cross-validation estimate of expected excess error, "... the difference in observed error when we don't or do let $\underline{x}_i^{T}$ assist in its own prediction" (Efron, 1982).

The sample estimate of $X^{T}X$ obtained upon deletion of the ith observation is

$$X^T_{(i)}X_{(i)} = X^TX - \underline{x}_i\underline{x}_i^T, \ \det\{X_{(i)}{}^TX_{(i)}\} = \det(X^TX\}(1-h_i)$$

And the cognate change in $\mathrm{Var}(\underline{\hat{\beta}})$ is described in terms of the ratio of the determinants, det{.}, of the respective matrices (Belsley et al, 1980):

$$\mathrm{COVRATIO} = \det\{\mathrm{Var}(\underline{\hat{\beta}}_{(i)})\}/\det(\mathrm{Var}(\underline{\hat{\beta}})\} = \left[\frac{n\text{-}k + e_i^{\#2}}{n\text{-}k\text{-}1}\right]^{-(k-1)} \frac{1}{(1-h_i)}$$

where $p = k-1$ and where $e_i^{\#} = e_i/\hat{\sigma}_{(i)}(1-h_i)$ is a Studentized residual.

"Large" values of COVRATIO identify those observations whose deletion strongly affects the covariance matrix, $\mathrm{Var}(\underline{\hat{\beta}})$, for a model. "Large" values are those for which $|\ \mathrm{COVRATIO}\text{-}1\ | > 3k/n$. "A value of COVRATIO greater than one indicates that the absence of the associated observation impairs efficiency, while a value less than one indicates the reverse" (Belsley et al, 1980). N.B.: "... some aspects of the stability of an estimate can be judged empirically by examining how much the estimate changes as observations are removed." (Cox and Hinkley, 1974). That is, perturbations of the sample such as represented by the deletion of individual data points will have the least effect on the most adequate model of the sample - which confirms its goodness-of-fit on yet another criterion. That is, $\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}}$, and $\mathrm{Var}(\underline{\hat{\beta}}_{(i)}) - \mathrm{Var}(\underline{\hat{\beta}})$, $1 \leq i \leq n$, for an adequate model; i.e., a statistically adequate model of a given sample is robust.

The statistics such as $\mathrm{RSS} = \Sigma e_i^2$, $\underline{\hat{\beta}}$, and $\mathrm{PRESS} = \Sigma e_{(i)}^2$ are aggregate statistics; i.e., the statistic takes only a single value for the n observations that comprise the sample. The statistics e_i, h_i, $e_{(i)}$, $\underline{\hat{\beta}}_{(i)}$, $\mathrm{Var}(\underline{\hat{\beta}}_{(i)})$ are case statistics; i.e., they may take a different value for each of the n observations, or cases, in the sample. In order to adequately assess the "fit" of any model to the data it is absolutely necessary not only to examine the aggregate measures such as the RSS, or functions thereof, in the context of their respective sampling distributions as in a standard goodness-of-fit test but also to examine the plots (distributions) of the case statistics; the latter are often - and aptly - referred to as regression diagnostics since they can identify those observations that are not well-explained by the model (e.g., observations at which e_i is "large") and those which may dominate one or more aspects of the "fit" (e.g., observations at which h_i, $\underline{\hat{\beta}}_{(i)}$ or $\mathrm{Var}(\underline{\hat{\beta}}_{(i)})$ are "large"). The residuals, e_i, and hat matrix diagonals, h_i, are key diagnostics since many of the others, e.g., $\underline{\hat{\beta}}_{(i)}$ are functions of these two. The set of e_i provide a measure of how well the model represents the sample. The set of h_i provide a measure of how well the model transcends the sample, i.e., how well it can be generalized. (A model that holds for only one set of data represents merely a unique historical artifact - or a cultural event. Ehrenberg, 1975.)

These case statistics can quantify both the degree of non-uniformity that may be present in the data, and its specific effects on the estimates and inferences made therefrom. The investigator must recognize that the naive application of ordinary least squares analysis to "fit" a given model to a set of data is not a robust procedure. It is often the case that only modest perturbations of the data (such as omission of an observation) can result in substantial changes in the estimates, say large values of $(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)})$, and inferences, say on the concordance, obtained from the fitted model. However, if the naive least squares analysis is augmented by the appropriate regression diagnostics, then the entire estimation procedure becomes more robust, since the outlying and/or influential observations are identified and their effects on the estimates and inferences can be assessed and, if required, appropriate corrective measures can be taken; for example, down-weighting, or even deletion, of observations; transformation, say $x \longrightarrow \log x$, of one or more variables, etc.

N.B.: Although the residuals, e_i, and hat matrix diagonals, h_i, are key diagnostics there are some problems attending their deployment that are due in part to the fact that each is a function of the mean vector and covariance matrix of the sample, neither of which are robust: a small cluster of outlying observations will attract the mean vector and inflate the variance-covariance matrix. The principal difficulties have been described and a solution proposed by Rousseeuw and van Zomeren (1990): "Detecting outliers in a multi-variate point cloud is not trivial, especially when there are several outliers. The classical identification method does not always find them, because it is based on the sample mean and covariance matrix, which are themselves affected by the outliers. That is how the outliers get masked. To avoid the masking effect, we propose to compute distances based on very robust estimates of location and covariance. These robust distances are better suited to expose the outliers.

In the case of regression data, the classical least squares approach makes outliers in a similar way. Also here, the outliers may be unmasked by using a highly robust regression method. ... a new display is proposed in which the robust regression residuals are plotted versus robust distances. This plot classifies the data into regular observations, vertical outliers, good leverage points, and bad leverage points." (N.B.: "regular observations" are defined by the majority of the data; a "good leverage point" is an outlier that increases the precision of the parameter estimates; a "bad leverage point" is an outlier that decreases the precision of the parameter estimates; a "vertical outlier" is an outlier that is neither a good nor bad leverage point.)

One especially useful transformation of predictor variables seems to be frequently overlooked in the received models of dose-response. This is the spline transformation (Montgomery and Peck, 1982):

$$x \longrightarrow (x-x_0)_+ = \begin{cases} (x-x_0), & x \geq x_0 \\ 0, & x < x_0 \end{cases}$$

or

$$\log x \longrightarrow \log(x/x_0)_+ = \begin{cases} \log(x/x_0), & x \geq x_0 \\ 0, & x < x_0 \end{cases}$$

These $(.)_+$ functions represent the linear splines in which x_0 is the so-called "knot". The splines may be used to represent a "threshold" in the dose-response process that generated the observations. Estimates of x_0 may be obtained either a priori or from the sample (by an iterative procedure. See Herbert, 1986b). Examples of generalized linear model processes with thresholds in dose (Poisson response) and in time (Binomial response) are presented in Herbert (1986b) and Herbert (1985a), respectively.

We have reminded the reader several times that least squares estimates of the parameter vector, $\underline{\beta}$, are not robust in that a single, usually outlying, observation can strongly influence the estimates $\hat{\underline{\beta}}$, Var($\hat{\underline{\beta}}$) and the goodness-of-fit measure. Another alternative estimation procedure that is also more robust than "naive least squares" for the Normal theory model is one that provides estimates of the parameter $\underline{\beta}$ which minimize the sum of absolute residuals, $\Sigma|e_i|$, rather than the sum of squared residuals, Σe_i^2 (See Montgomery and Peck, 1982). The absolute error criterion weights outlying observations far less than does the least squares criterion. However, it must be formulated as a linear programming problem and hence requires an iterative solution (e.g., SIMPLEX). Moreover, the estimates of $\underline{\beta}$ so obtained are not unbiased, whereas as the parameter estimates obtained by ordinary least squares methods are.

Still another simple robust procedure, based on the weighted jackknife pseudovalues, $Q_i = \hat{\underline{\beta}} + n(1-h_i)(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})$, has been described by Hinkley (1974). Note that the Q_i are weighted functions of the row-deleted estimates, $\hat{\underline{\beta}}_{(i)}$, where the weights are $(1-h_i)$ and h_i is the hat matrix diagonal. Hinkley remarks that in his procedure, unlike other robust methods such as those of Huber, "... 'harmless' large residuals are ignored" since Q_i will not be extreme unless h_i is large.

Robust estimation procedures based on weighted jackknife methods are readily generalized to non-Normal theory models, whereas those based on the least absolute deviations methods are not so readily generalized.

We had remarked above that although the sample estimate, $\hat{\beta}_j$, of the weight, β_j, to be attached to the j^{th} variable, X_j, $1 \leq j \leq p$, of a multivariate regression model depends upon the relation, as described by the correlation coefficient, r_{jk}, of the sample observations on X_j to those of the other p-1 variables, a realizing sense of this contingency is not evident in much of the published literature. We have referred to this as one of ontological weaknesses in the current praxis of radiobiology. It has an important epistemilogical correlate.

A vivid appreciation of the situation can be achieved by considering the simplest multivariate model, a bivariate Normal theory model, p=2, in which the columns in the observation matrix [$\underline{y}$, X] have been standardized to unit length (Montgomery and Peck, 1982):

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.\ 1 \leq i \leq n.$$

The least squares estimates of the parameters β_1 and β_2 are

$$\hat{\beta}_1 = (r_{1y} - r_{12}r_{2y})/(1 - r_{12}^2)$$
$$\hat{\beta}_2 = (r_{2y} - r_{12}r_{1y})/(1 - r_{12}^2)$$

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2/(1 - r_{12}^2),\ 1 \leq j \leq 2$$

$$\mathrm{Cov}(\hat{\beta}_1,\ \hat{\beta}_2) = -r_{12}\sigma^2/(1 - r_{12}^2)$$

Here r_{jy} is the correlation between $\underline{y}$ and X_j and r_{12} is the correlation between X_1 and X_2.

It is evident at once that the respective parameter estimates, $\hat{\beta}_j$, as well as $\mathrm{Var}(\hat{\beta}_j)$, j = 1,2, and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ are strong functions of the relation of X_j to the other variable - a relation that is described by r_{12}. Large values of r_{12} - a condition referred to as multicollinearity - will inflate $\hat{\beta}_j$, $\mathrm{Var}(\hat{\beta}_j)$, etc.; it may also result in one or more of the $\hat{\beta}_j$ having the wrong sign. The inflation of $\mathrm{Var}(\hat{\beta}_j)$ will cause the point estimates, $\hat{\beta}_j$, to be unstable.

Figures 3 and 4 provide examples of dose-response data in which two variables, N and T, are highly correlated. In general, N and T will always be highly correlated in samples of clinical data and hence the sample estimates of the respective effects of each of these covariates on the levels of response will be quite unstable. It is important to note as well that unless the range of dose, D, in a given sample is quite large, the variables D and D^2 are highly correlated and hence the point estimates of the respective coefficients α and β in the so-called LQ model will be unstable and may even have the wrong sign.

Measures of the degree of multicollinearity present in the sample distribution of the observations in X are provided by still other diagnostics. One measure in the condition number $\kappa = \lambda(p)/\lambda(1)$ where $\lambda(p)$ is the largest and $\lambda(1)$ is the smallest of the eigenvalues of the p*p correlation matrix $C = X_0^T X_0$ where the subscripts denote that each of the predictor variables have been standardized. Values of $\kappa > 100$ suggest that multicollinearity may be a problem. See Montgomery and Peck (1982).

A measure of one of the more important effects of the presence of multicollinearity, that is, inflation of $\mathrm{Var}(\hat{\underline{\beta}})$, is provided by the variance inflation factor, VIF_j, $1 \leq j \leq p$. This diagnostic measures the inflation of $\mathrm{Var}(\hat{\beta}_j)$ with respect to the variance for uncorrelated predictor variables. $\mathrm{VIF}_j = c^{jj}$, the j^{th} diagonal element of C^{-1}, the inverse of the correlation matrix, $C = X_0^T X_0$. Values of $\mathrm{VIF}_j > 10$ suggest that multicollinearity may be a problem. See Montgomery and Peck (1982).

It can be shown that the VIF_j are functions of the eigenvalues λ_j and eigenvectors V_j (with components V_{ij}), of the correlation matrix, C:

$$\mathrm{VIF}_j = \sum^{p} v_{ji}^2/\lambda_i$$

The largest components of the eigenvector associated with the smallest eigenvalue identify those predictor variables, say X_j, that are responsible for the greater part of the multicollinearity. Although there is no formal criterion for identifying "large" components, most investigators consider any component whose absolute value exceeds 0.1 to be "large".

It is important to remark that the diagnostics, such as e_i (and e_i^*)m g_i, $\hat{\underline{\beta}}_{(i)}$, κ, etc., are, for a given set of data, strongly model-dependent. That is, these diagnostics identify model-sample interaction. Changes in the range of observations on a given variable, changes in the transform of a given variable (e.g., $x_i \longrightarrow \log x_i$), deletion of one (or more) observations or one (or more) variables, changes in the link function of the model (e.g., from identity to probit) may materially - and significantly - change the estimates and inferences obtained from a study. We present an abundance of examples of each of these model-sample interactions in the sections to follow as well as in the Annexes.

It should be noted that the set of n diagonal elements, h_i, $1 \leq i \leq n$, of the hat matrix, H, and the set of p eigenvalues, λ_j, $1 \leq j \leq p$, of the correlation matrix, C, provide very useful criteria for the evaluation of experimental designs for estimation of the parameters of generalized linear models of dose-response: For an adequate design $h_i - ... - h_n - p/n$ and $\lambda_1 - .... - \lambda_p -$ 1.0. Or, "To minimize the effects of a small proportion of outlying responses on the fitted values, choose a design to minimize the dispersion of the $[h_i]$" (Cook and Weisberg, 1982). And, "A design is said to be E-optimal if it minimizes the maximal eigenvalue of $(X^TX)^{-1}$" (Steinberg and Hunter, 1984). See also section 7.11.1, below.

7.1.2 Non-parametric regression

Non-parametric regression is a set of smoothing procedures for estimating a regression curve without making strong prior assumptions about the shape of the true regression function, that is, when there is little a priori information on its shape. If parametric regression can be viewed (and it can) as a procedure for combining information across (predictor) variables then non-parametric regression can be viewed as a procedure for combining information across observations. It is, in fact, a method for "borrowing strength" across observations. Therefore, these procedures are useful for the construction and testing of parametric models -including residual analysis - as well as for the description of samples. For such procedures, a scalar parameter, say f, $(0 < f \leq 1)$ called the smoothing parameter, determines the number of neighboring observations, (y_i, x_i), that contribute to a weighted estimate, $\hat{y}_k$ at x_k, and hence determines the smoothness of the resultant regression curve. In the non-parametric regressions described in this report we have used only Cleveland's 1979 robust locally weighted regression (LOWESS) algorithm, but other smoothing algorithms give qualitatively similar results: "Robust locally weighted regression is a method for smoothing a scatterplot (x_i, y_i), $i = 1, ..., n$, in which the fitted value at x_k is the value of a polynomial fit to the data using weighted least squares, where the weight for (x_i, y_i), is large if x_i is close to x_k and small if it is not. A robust fitting procedure is used that guards against deviant points distorting the smoothed points." ... "The smoothing procedure has been designed to accommodate data for which

$$y_i = g(x_i) + \epsilon_i$$

where g is a smooth function and the ϵ_i are random variables with mean 0 and constant scale. ... The assumption of smoothness allows points in the neighborhood of (x_i, y_i) to be used in forming $\hat{y}_i$. ... Thus, points whose abscissas are close to x_i play a large role in the determination of $\hat{y}_i$ while points far away play a lesser role. Increasing f increases the neighborhood of influential points and therefore tends to increase the smoothness of the smoothed points. ... increasing f tends to increase the smoothness of the smoothed points $(x_i, \hat{y}_i)$" (Cleveland, 1979).

Proper selection of smoothing parameters is critical to the success of all smoothing algorithms: "The goal in the choice of f is to pick a value as large as possible to minimize the variability in the smoothed points without distorting the pattern in the data." (Cleveland, 1979). However, the choice of the smoothing parameter obviously also depends strongly on prior beliefs about the nature of the process generating the sample observations - in addition to it being described by a smooth function. Thus, although the use of non-parametric regression does indeed, "let the data speak," in the present case it is necessary for the investigator to select, a priori, the fraction, say f, of the data that speaks to any one point, say (y_k, x_k), $1 \leq k \leq n$, at a time. (The situation is quite analogous to Cluster Analysis in which the investigator must select, a priori, the number of clusters into which the data is to be partitioned by the clustering algorithm - say by k-means. See, for instance, Anderberg, 1973.)

7.2 Post-hoc salvage methods for parametric regression. "Regression therapeutics."

"We argue that it is typically true that there is available prior information about the parameters and that this may be exploited to find improved, and sometimes substantially improved, estimates."

D. Lindley and A. Smith, 1971

"Bayesian analysis is a method by which information in a given data set can be combined with other relevant evidence."

...

"All non-experimental data are simply too weak to allow sensible inferences in the absence of supplementary information."

E. Leamer, 1986

"... the factors which are considered biologically plausible have been promulgated in the light of current knowledge ... A lack of biological plausibility may indicate the limitations of our knowledge rather than the real lack of a causal association ... biological plausibility is the weakest kind of evidence for assessing cause-effect relationships."

E. Neugebauer et al, 1987

"It could be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship ta scribe the typhus which he there contracted, to the vermin with which the bodies of the sick might be infected. An adequate cause, one reasonable in itself, must correct the coincidence of simple experience."

D. W. Cheever, 1861

For the Normal theory model $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, $\underline{\varepsilon} - N(\underline{0}, \sigma^2 I_n)$ of the observation matrix [ym X] we have $\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$, $Var(\hat{\underline{\beta}}) = \sigma^2(X^TX)^{-1}$, $RSS = (\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\beta}) = \Sigma e_i^2$. The sample information on $\underline{\beta}$ includes the observation matrix [$\underline{y}$, X], and the sample estimates $\hat{\underline{\beta}}$ and $Var(\hat{\underline{\beta}})$. We now consider several procedures by which such information may be <u>augmented</u> by non-sample, or a priori, information. In each procedure a reduced-variance estimator of $\underline{\beta}$ is obtained as a matrix-weighted average of sample and non-sample information on $\underline{\beta}$. As we remarked above, we shall be especially interested in prior information that can be represented in a form <u>parallel</u>, or isomorphic, to that of the regression model $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$. That is, we are interested in prior information that can be expressed in <u>matrix form</u>.

7.2.1 <u>Bayes Estimation</u>

We assume first that the <u>a priori</u> information on the parameter vector $\underline{\beta}$ can be described by a multivariate Normal distribution with mean $\underline{\beta}_0$ and covariance matrix Σ_0. Then the Bayes estimate of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}}^{**} = [\phi X^TX + \Sigma_0^{-1}]^{-1}(\phi X^T\underline{y} + \Sigma_0^{-1}\hat{\underline{\beta}}_0) = [\phi X^TX + \Sigma_0^{-1}]^{-1}(\phi X^TX\hat{\underline{\beta}} + \Sigma_0^{-1}\hat{\underline{\beta}}_0)$$

and

$$Var(\hat{\underline{\beta}}^{**}) = [\phi X^TX + \Sigma_0^{-1}]^{-1} \text{ where } \phi = 1/\hat{\sigma}^2.$$

Obviously, the Bayes estimate is a matrix-weighted average of a priori ($\underline{\beta}_0$, Σ_0) and sample ($\hat{\underline{\beta}}$, ϕX^TX) information on $\underline{\beta}$.

7.2.2 <u>Ridge Regression</u>

A statistical procedure that provides both a measure of the effect - inflation of $\hat{\underline{\beta}}$ and $Var(\hat{\underline{\beta}})$ - of the presence of collinearity in the observation matrix [$\underline{y}$, X], that is, a <u>regression diagnostic</u>, that is complementary to the variance-inflation factor, VIF_j, and also a technique for ameliorating this effect - producing a "shrunken estimator" of $\underline{\beta}$ - is the procedure of Ridge regression. The Ridge estimator, $\hat{\underline{\beta}}_R$, is formally equivalent to a matrix-weighted average of sample and prior information in which the prior estimate of $\underline{\beta}$ is $\hat{\underline{\beta}} \equiv \underline{0}$, the null vector and the covariance matrix, kI_p, where $o \leq k \leq \infty$ and I_p is the p*p identity matrix. k, the so-called biasing parameter is determined by the investigator on the basis of several criteria to be described below (Leamer, 1978). The Ridge procedure is used when the a priori information on $\underline{\beta}$ is very <u>weak</u>, that is, no more than the respective sign and significance of each of the components, $\bar{\beta}_j$, $1 \leq j \leq p$, of $\underline{\beta}$ is known, a priori, i.e., $\beta_1 < 0$, $\beta_2 > 0$, etc.

Ridge regression was initially developed for Normal theory models (Hoerl and Kennard, 1970. See also Marquardt and Snee, 1975). Schaefer (1984) and Schaefer et al (1984) have developed Ridge regression methods for the logistic model of binary response data. However, there has been little development in logistic Ridge regression since then although there have, of course, been several hundred papers dealing with Ridge regression for Normal-theory models published.

We illustrate the procedure for the Normal model, $\underline{y} = X\hat{\underline{\beta}} + \underline{\varepsilon}$, now written in the <u>correlation basis</u>. The observation matrix [$\underline{y}$, X] has been centered and scaled so that $X^T\underline{y}$ and X^TX are <u>correlation matrices</u>. The LS estimator is $\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$. The Ridge estimator is $\hat{\underline{\beta}}_R = (X^TX + kI_p)^{-1}X^T\underline{y}$ where $0 \leq k < \infty$ is the <u>biasing parameter</u> and I_p is the (p*p) identity matrix. For k

$= 0$, $\hat{\underline{\beta}}_R = \hat{\underline{\beta}}$, i.e., the Ridge and LS estimators coincide. As $k \longrightarrow \infty$, $\hat{\underline{\beta}}_R \longrightarrow [0]$, the null vector. The Ridge estimator is a linear transform of the LS estimator:

$$\hat{\underline{\beta}}_R = (X^T X + kI_p)^{-1} X^T X \hat{\underline{\beta}}$$

and

$$Var(\hat{\underline{\beta}}_R) = (X^T X + kI_p)^{-1} Var(\hat{\underline{\beta}})(X^T X + kI_p)^{-1}.$$

$$RSS_R = (\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}}) + (\hat{\underline{\beta}}_R - \hat{\underline{\beta}})^T X^T X(\hat{\underline{\beta}}_R - \hat{\underline{\beta}})$$

where $RSS = (\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})$. Thus, the Ridge estimator, $\hat{\underline{\beta}}_R$, does not fit the sample data as well as the least-squares estimator, $\hat{\underline{\beta}}$. That is, the Ridge estimator is a non-least squares estimator. (Weisberg, 1985). However, it is usually the case that for Normal theory models of data for which the degree of collinearity is quite high, the Ridge estimator of $\underline{\beta}$ - as a shrunken estimator - will provide a better fit in extrapolation than does the least squares estimator, $\hat{\underline{\beta}}$ (vide infra).

If a high degree of collinearity is present in the matrix X, then for small values of k, say $0 < k < 0.05$, the difference $(\hat{\underline{\beta}} - \hat{\underline{\beta}}_R)$ is large.Therefore, the Ridge procedure provides a kind of sensitivity test of the LS sample estimate $\underline{\beta}$. Note that this may be regarded as a column diagnostic similar to row diagnostics, such as Cook's D_i. Each is a description of the behaviour of $\hat{\underline{\beta}}$ under small perturbations of the data: 1) the matrix kI_p of Ridge regression and 2) the column vector, $\phi\underline{d}$, of the mean shift outlier model, respectively.

The effects of the small perturbations, kI_p, where I_p is a (p*p) identity matrix on the sample estimate of the parameter vector $\underline{\beta}$, which disclose the degree of collinearity in X, are most vividly displayed in the so-called Ridge trace which is a plot of the components of $\hat{\underline{\beta}}_R$ vs k. The Ridge trace can also be used to select a value of k at which $\hat{\underline{\beta}}_R$ is sufficiently stable to provide a useful estimator of $\underline{\beta}$ (when re-transformed from the correlation basis to the observation basis, i.e., in which $\hat{\sigma}^2 X^T X$ is a covariance matrix). On this criterion the optimal degree of "shrinkage" of the OLS estimate of $\underline{\beta}$ is obtained at that value of k for which the Ridge traces for all of the p coefficients are nearly "flat".

There are several other criteria for selecting an optimal value of the biasing parameter k. (See for instance, Montgomery and Peck, 1982; Smith and Campbell, 1980.) We mention only three. a) The Marquardt and Snee (1975) criterion: Choose k so that $VIF_j(k) - 1.0$, $0 \leq j \leq p$. b) The Obenchain (1977) criterion: Choose k so that the Ridge estimate, $\hat{\underline{\beta}}_R$, lies within the 0.90 confidence ellipsoid on $\underline{\beta}$. c) The Hoerl, Kennard, and Baldwin (1975) criterion: $k = p\hat{\sigma}^2/\hat{\underline{\beta}}^T\hat{\underline{\beta}}$ where $p = k - 1$ the number of predictor variables.

The Ordinary Least Squares (OLS) estimate, $\hat{\underline{\beta}}$, maximizes the correlation between the observed $\underline{y}_I$ and estimated, $X_I\hat{\underline{\beta}}$, responses in the original data $[\underline{y}_I, X_I]$ from which $\hat{\underline{\beta}}$ was obtained. The Ridge regression (RR) estimate $\hat{\underline{\beta}}_R$ maximizes the correlation between the observed, $\underline{y}_{II}$, and expected, $X_{II}\hat{\underline{\beta}}_R$, responses in new data, $[\underline{y}_{II}, X_{II}]$. Thus, for highly collinear data in which $(\hat{\underline{\beta}} - \hat{\underline{\beta}}_R)$ is large, the RR estimate, $\hat{\underline{\beta}}_R$, gives better predictions in extrapolation than does $\hat{\underline{\beta}}$.

Non-least squares estimators of model parameters such as RR belong to the broad class of shrunken estimators. (Berger, 1983; Casella, 1985; Efron and Morris, 1973). The shrunken estimators offer improvements - in some sense, usually mean-squared error, and in some cases - on the usual least squares (LS) and maximum likelihood (ML) estimators. It should be noted that the predictive performance of a given model in extrapolation (i.e., in "new data") should be better for the shrunken estimator of the parameter vector than for the LS estimator since the effects (inflation of $\hat{\underline{\beta}}$ and $Var(\hat{\underline{\beta}})$) of multicollinearity - when present - in the construction sample data are reduced by post-hoc shrinkage procedures. Let us consider the simplest example of a shrunken estimator. For instance, in obtaining an estimate of an unknown parameter θ it may be that the investigator believes that there is good a priori information that the value of the parameter takes a specific value, say θ_0. In such cases, it may be reasonable to "shrink" the usual estimator, say $\hat{\theta}$, toward θ_0 by adding a fraction, k, of the difference $(\hat{\theta} - \theta_0)$ to θ_0 to give the shrunken estimate, $\hat{\theta}_s = k(\hat{\theta} - \theta_0) + \theta_0 = k\hat{\theta} + (1 - k)\theta_0$, $0 \leq k \leq 1$. $(1 - k)$ is directly proportional to the strength of the investigator's belief in θ_0. The resulting estimator, though perhaps biased, may have a smaller mean square error than $\hat{\theta}$ for θ in some interval around θ_0. In practice, one often tests the hypothesis H_0: $\theta = \theta_0$. If θ_0 is not rejected then one uses the shrinkage estimator $\hat{\theta}_s$ for θ;

if H_0 is rejected one uses $\hat{\theta}$ (Lemmer, 1983). The class of shrunken estimators includes the so-called Stein-type estimators as well as Ridge estimators. Both the RR and Stein-type estimators can be easily formulated as linear transforms of the LS estimator.

The James-Stein estimator, $\hat{\underline{\beta}}^{***} = [1 - (k-3)\hat{\sigma}^2/(\hat{\underline{\beta}}^T X^T X \bar{\underline{\beta}}]\hat{\underline{\beta}}$, is included for comparison with the Ridge estimator. For $p \geq 3$ it has been shown that this shrunken estimator corrects for the over-fitting that occurs in Normal theory multiple regression and always has lower expected error for predicting $\underline{y}$ than does the LS estimator (Darlington, 1978). In both James-Stein and Ridge regression the LS estimator is shrunk toward the origin: $\underline{\beta} = \underline{0}$. "Ridge estimators were designed as a method to improve on the unsatisfactory characteristics of the least-squares estimator when there is multicollinearity present - when X^TX is badly conditioned. Stein-type estimators are frequently recommended because they reduce mean-square error and they can be regarded as empirical Bayes estimators" (Rolph, 1976). Although, "Both Stein-type regression and ridge regression are empirical Bayes procedures. They are Bayes procedures because they combine prior and empirical estimates of parameters. They are empirical Bayes procedures because the data rather than the prior confidence statements determine the relative weights of the prior and empirical estimates. This avoids one of the worst pitfalls of pure Bayes procedures - seriously wrong prior estimates in which the investigator has great confidence so that they are little changed even by large amounts of data." (Darlington, 1978).

In general, the Stein, as well as the RR estimators, will be more stable and will, on the average, be closer to the true parameter values. Moreover, because of the latter property, their predictive performance in extrapolation to new data will be better than their unmodified least squares and/or maximum likelihood counterparts. Both the shrunken estimators and the Mixed estimators described below are reduced-variance regression (RVR) estimators. There is more to say about Stein estimators in the paper "Overview of Some Concepts, Methods, and Criteria New to Radiobiological Modelling" included in the Proceedings of the 4th ICDTF (Herbert, 1993b).

Ridge regression is only one of several salvage operations that can be deployed, post hoc, to retrieve information on a regression model from a study in which the data are weakened by the presence of multi-collinearity in the distribution of the observations in the sample, $[\underline{y}, X]$. The other salvage maneuvers are Data Augmentation and Mixed Estimation. These latter two maneuvers can be deployed to salvage studies encumbered by other weaknesses in the data, $[\underline{y}, X]$, in addition to multicollinearity, e.g., non-uniformity. We discuss Data augmentation first.

7.2.3 Data augmentation

Pooling the data of several different weak studies of the same hypothesis is the earliest form of meta-analysis (Glass et al, 1981). The maneuver of pooling the original $(\underline{y}_1, X_1)$ collinear data with the additional data chosen to reduce the effects of the presence of collinearity, influential observations, etc., in $[\underline{y}, X]$, on the sample estimates of $\underline{\beta}$ may be represented by an equation in the combined observation matrices as follows: (See Theil 1971 and Johnston 1972).

$$\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \underline{\beta} + \begin{pmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{pmatrix}$$

Then the unbiased estimate of $\underline{\beta}$ is

$$\hat{\underline{\beta}}^* = [(1/\sigma_1{}^2 X_1{}^T X_1 + (1/\sigma_2{}^2 X_2{}^T X_2]^{-1}[X_1{}^T \underline{y}_1 + X_2{}^T \underline{y}_2]$$

and

$$Var(\hat{\underline{\beta}}^*) = [(1/\sigma_1{}^2)X_1{}^T X_1 + (1/\sigma_2{}^2)X_2{}^T X_2]^{-1}$$

That is, $\bar{\underline{\beta}}^*$ is a matrix-weighted average of the initial and augmenting data.

Dykstra (1971) and Gaylord and Merrill (1968) describe methods by which the augmenting data $[\underline{y}_2, X_2]$ can be systematically selected to give an optimum distribution of $[\underline{y}, X] = [\underline{y}_1, X_1] \oplus [\underline{y}_2, X_2]$.The criterion is to choose $[\underline{y}_2, X_2]$ to maximize the determinant $|X^TX|$. This minimizes both the volume of the confidence region for the regression coefficients and the maximum variance of a predicted value, $Var(\hat{y}_i)$. The extremes are achieved by obtaining responses at the extremes of the region of observations. For a first-order model in p treatment variables these are the observations at the 2p corners of the polytope. These methods may salvage "... undesigned, nonorthogonal experimental data in those cases in which it is desired to separately estimate and test the effects of the independent variables without discarding the existing data. Moreover, these methods provide ... better estimates of the regression coefficients than discarding the original data and conducting an orthogonal experiment ..." (Gaylord and Merrill, 1968) with the same number of data points.

We have described the maneuver of data-augmentation in terms of the Normal theory model; it may be readily generalized in obvious ways to data in which the response has either a Binomial or Poisson distribution. In this report we have described reported instances of its deployment in the case of Poisson response for the LSS data on leukemia incidence and the Sparrow et al data on mutagenesis. In each case the data that were pooled, $[\underline{y}_1, X_1]$, $[\underline{y}_2, X_2]$, ..., had complementary weaknesses. Note, however, that in each of these studies, the data augmentation maneuver was not deployed in any formal way such as described by Gaylord and Merrill (1968). Nor, for that matter, in either study was there any acknowledgment that any salvage maneuver was being carried out - or was even made necessary by the characteristic weaknesses in the respective samples. See Annex III, parts 4 and 5 of this report.

7.2.4 Mixed estimation. Sensitivity analysis.

"No matter how strong be our degree of belief, we must always bear in mind that empirical evidence is never complete."

W. Edwards Deming, 1986

"It is my impression that rather generally, ..., it is considered decent to use judgment in choosing a function form but indecent to use judgment in choosing a coefficient. If judgment about important things is quite alright, why should it not be used for less important ones as well?"

J. Tukey, 1976

"Nevertheless, we must recognize that with weak data we have little choice but to incorporate prior beliefs into the analysis ..."

"As a first approximation, weighted averages of one's prior beliefs and data information would often seem preferable to us, although proper combination of prior information and data information involves more complex procedures."

Robins and Greenland, 1986

Robins and Greenland (1986) - and, of course, others - have remarked, that in model selection it is sound strategy to "... choose a model that is consistent with the data and yields parameter estimates consistent with ... prior beliefs." Or, as Sir Arthur Eddington (1959) has recommended, in what some might regard as an instance of flaming recidivism, do not, "put overmuch confidence in the observational results until they have been confirmed by the theory." However, common practice, as disclosed by the journal literature on radiation dose-response models ignores both recommendations: the consistency of the model with the data - measured by the "goodness-of-fit" - is rarely assessed by statistically adequate measures and the consistency of the sample estimates, $\hat{\underline{\beta}}$, and $\underline{x}^T\hat{\underline{\beta}}$, with prior beliefs is never adequately examined at all, although the prior beliefs, variously implemented, are regularly deployed in received praxis to achieve the estimates required by the investigators. (As examples of recent unexamined deployments of a priori information see BEIR III (1980) pp 187-188 and Tables V-9 and V-10, Tucker and Thames 1983, p 1378 and Fig. 8. See also Fig. 11c this report.) Mah et al 1987, Travis and Tucker 1987, and Fowler 1991.

As remarked above, matrix theory, which we introduced to describe the multivariate nature of the response and the subsequent construction and assessment of a multivariate model, $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, of the sample data $[\underline{y}, X]$ also provides an isomorphic representation for non-sample information that is required for 1) combining, in the posterior estimate $\bar{\underline{\beta}}^{**}$, the sample information on the parameter vector $\underline{\beta}$ of the model, with a priori, non-sample, information obtained from theory, previous experiments, introspection, etc.; 2) for testing, by a statistic γ, the consistency of prior and sample information; and 3) for determining the proportion, θ_p, of the a priori information on $\underline{\beta}$ that is included in $\underline{\hat{\beta}}^{**}$.

Mixed estimation uses a prior i information on $\underline{\beta}$ to <u>augment</u> the sample data, [$\underline{y}$, X], directly instead of through a prior distribution as in Bayesian regression (Montgomery and Peck, 1982). Mixed estimation is useful when the a priori information on $\underline{\beta}$ that is available is <u>more precise</u> than the prior information which can be deployed in the post hoc maneuver of Ridge regression. That is, either the values of q of the parameters are known, or the values of functions, e.g., ratios, of q of the parameters are known, a priori. Here $1 \leq q \leq p$ where $p = k + 1$ is the number of freely adjustable parameters in the model with a (k*1) parameter vector $\underline{\beta}$.

The sample information on the model is represented by the estimates $\underline{\hat{\beta}}$ and $\mathrm{Var}(\underline{\hat{\beta}})$. The prior information is represented by the linear constraint $\underline{r} = R\underline{\beta} + \underline{v}$, where $\underline{r}$ and $\underline{v}$ are (qx1) matrices, R is (q*k) and $\underline{\beta}$ is (k*1); $E(\underline{v}) = \underline{0}$ where $\underline{0}$ is the null vector and $\mathrm{Var}(\underline{v}) = \psi$. ψ is a measure of the precision of the a priori information. It may also represent the "strength of belief" in the a priori information. If the elements of the matrices $\underline{r}$ and R of the a priori constraint on $\underline{\hat{\beta}}$ are taken from the sample estimate of $\underline{\beta}$ obtained in a prior study with covariance matrix $\mathrm{Var}^*(\underline{\hat{\beta}})$ then $\psi = R\mathrm{Var}^*(\underline{\hat{\beta}})R^T$.

It is useful to represent the combination of the two sources of information on the parameter vector $\underline{\beta}$ as the system of matrix equations:

$$\begin{pmatrix} \underline{y} \\ \underline{r} \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \underline{\beta} + \begin{pmatrix} \underline{\varepsilon} \\ \underline{v} \end{pmatrix}$$

Note that this implementation of the prior information [$\underline{r}$, R] on $\underline{\beta}$ is identical to that for data augmentation, [$\underline{y}_2$, X]. Application of least squares methods to this system of matrix equations gives the mixed, or posterior, estimate of $\underline{\beta}$ as the matrix-weighted average

$$\underline{\hat{\beta}}^{**} = [\phi^{-1}X^TX + R^T\psi^{-1}R]^{-1}(\psi^{-1}X^T\underline{y} + R^T\psi^{-1}\underline{r})$$

and

$$\mathrm{Var}(\underline{\hat{\beta}}^{**}) = [\psi^{-1}X^TX + R^T\psi^{-1}R].$$

It can be shown that $\underline{\hat{\beta}}^{**}$ minimizes the weighted sum of squares, $\underline{\varepsilon}^T\underline{\varepsilon}/\sigma^2 + \underline{v}^T\psi^{-1}\underline{v}$. However, the sum of squared residuals for $\underline{\hat{\beta}}^{**}$ exceeds that for $\underline{\hat{\beta}}$:

$$RSS^{**} = RSS + (\underline{\hat{\beta}} - \underline{\hat{\beta}}^{**})X^TX(\underline{\hat{\beta}} - \underline{\hat{\beta}}^{**})$$

The <u>consistency</u> of the sample and <u>non-sample</u> information on $\underline{\beta}$ is assessed by the <u>compatibility</u> statistic γ which is distributed asymptotically as chi-squared on q degrees of freedom:

$$\gamma = (r - R\underline{\beta})^T[\phi R(X^TX)^{-1}R^T + \psi^{-1}](\underline{r} - R\underline{\beta}), \text{ where } \phi = \sigma^2.$$

The <u>proportion</u>, θ_p, of a priori information included in $\underline{\hat{\beta}}^{**}$ is given by

$$\theta_p = k^{-1}\mathrm{Trace}\{R^T\psi^{-1}R[\phi^{-1}X^TX + R^T\psi^{-1}R]^{-1}\}$$

where $0 \leqslant \theta_p \leqslant 1.0$. θ_p is a measure of the influence of the a priori information on the posterior estimate, $\underline{\hat{\beta}}^{**}$. Thus, it can be regarded as a "regression diagnostic" for the a priori information, rather analogous to Cook's distance, D_i, which is, of course, a (standardized) measure of the <u>influence</u>, $(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)})$, of the i^{th} sample observation on the sample estimate, $\underline{\hat{\beta}}$. Note that Mixed estimation and Ridge regression can be shown to be numerically equivalent. (Montgomery and Peck, 1982).

In any method that combines sample and a priori information on model parameters it is most important to recognize that the a priori, as well as sample, information is encumbered with uncertainty. In any <u>useful</u> specification of the a priori information this uncertainty <u>must</u> be described quantitatively. In the present case this uncertainty can be described by the (k*k) covariance matrix Σ_0 or the (q*q) covariance matrix ψ for the Bayesian and Mixed estimation methods, respectively. As Leamer (1986) has remarked, it is often more difficult to specify the elements of these covariance matrices, Σ_0 and ψ, than the elements of the vector, $\underline{\beta}_0$, or of the linear constraint, $\underline{r} = R\underline{\beta}$. That is, it is routinely much easier to state what we "know" than to say how well we know it. Nonetheless, it is also usually possible, with a little thought, to elicit a priori estimates both of the elements of $\underline{\beta}$ and of the covariance matrix, ψ. For instance, the investigator may know that a parameter, say β_1, almost certainly lies between 0 and 1 and very likely between 0.25 and 0.75. Then the constraint can be written as $0.5 = \beta_1 + v$, $E(v) = 0$, $\mathrm{Var}(v) = 0.0625$. Values of β_1 outside the range (0,1) would then be outside the "2-sigma" range: 0.50 ± 2*0.25.

(Theil and Goldberger, 1961). A priori estimates of the correlations, say Corr(β_1, β_j), of β_1 with other parameters, β_j, may also be introduced.

Note that the currently received procedures for combining a priori (animal) and sample (clinical) information on the parameter vector of dose-response models (the procedures described, for example, by Mah et al 1987, Travis and Tucker 1987, and Fowler 1991, provide examples of the received practice in this matter) do not provide for including any measures of the uncertainty that may encumber the a priori information. Thus, in current, received, practice it is possible for seriously wrong prior estimates to dominate conflicting sample estimates even when the latter are obtained from a large amount of excellent data.

Leamer, in addition, also recommends an analysis of the sensitivity of the posterior estimates of $\underline{\beta}$ to a range of a priori choices of the elements of $\underline{\beta}$, the constraint matrices ($\underline{r}$, R) and the covariance matrices Σ_0 and ψ. If the sensitivity analysis discloses that a "credibly narrow" range of values of Σ_0 or ψ will provide useful posterior estimates of $\underline{\beta}$ then the data set is acceptable. If, on the other hand, an "incredibly narrow" range of values of the covariance matrices is required to stabilize the sample estimates, $\hat{\underline{\beta}}$, then it can be concluded that the data at hand are not useful to the purposes for which they were acquired. They are unsalvageable. And note again that as the precision of the a priori information increases,the proportion of prior information included in the posterior estimates must also increase. (We have often found that the published values of the parameter vector for a given model of a given data set depend almost entirely on the a priori information incorporated therein, relegating much of the evidence of the data to a kind of talismanic role in the study: the data are invoked rather than "fit" to provide empirical support for the results. See Nisbett and Ross 1980, Chapt. 8. "theory maintenance and theory change.")

There are three distinct post-hoc salvage maneuvers in which either data, or statistics, or other information from two, or more, weak, parallel studies of a common issue are pooled: 1) Mixed or Bayesian estimation, including Ridge estimation, in which sample and non-sample information (estimates) on the parameter $\underline{\beta}$ of a regression model are pooled. 2) Data augmentation in which the data of two or more samples are pooled. 3) Classical meta-analysis, in which the inferential statistics (say, p-values) from several homogeneous studies are pooled, e.g., Stouffer's method for combining p-values from k different studies of the same null hypothesis: $z(k) = \Sigma z_i/\sqrt{k}$ where $z_i = \Phi^{-1}(p_i)$, $1 \leq i \leq k$, $\Phi(z_i)$ is the standard Normal distribution and p_i is the p-value for the i^{th} study. The null hypothesis is rejected whenever z_k exceeds the appropriate critical value of the standard Normal distribution. The respective z_i values of the k component studies can be weighted - either for importance or for precision. See Hedges and Olkin, 1985. We have described and instanced the use of the first two methods in this report.

The Bayesian hierarchical meta-analysis of DuMouchel and Harris (1983) to be described in section 14.3 below is a special case of Bayesian and/or Mixed estimation described above; it is especially useful in problems of inter-species transfer of dose-response functions ("mouse-to-man" extrapolation).

7.3. Sample estimates of $\theta = \beta_1/\beta_2$

The ratio of the coefficients of the linear quadratic model has been invested with considerably ontological significance. At issue is whether it should be estimated by the non-linear function, $\theta = \beta_1/\beta_2$, of the coefficients of the linear form $(\beta_1 D + \beta_2 D^2)$ or by the coefficient γ of the non-linear form $\beta_2(\gamma D + D^2)$. These are the indirect and direct estimates, respectively.

7.3.1 Indirect estimates of $\theta = \beta_1/\beta_2$. Weighted jackknife methods.

It is well-known that the ratio, $\theta = \beta_1/\beta_2$, of the coefficients of the linear and quadratic terms in dose in the LQ model has been invested with diverse biological meanings and usages. Therefore, the bias and variance of the sample estimate $\hat{\theta}$ are of no little interest. The ratio, $\theta = f(\underline{\beta})$ is a non-linear function of the parameter vector $\underline{\beta}$ of the LQ model. Unlike linear functions of the parameter vector such as (say) the response, $\underline{x}_i^T\underline{\beta}$, at the ith level of treatment variables, the maximum likelihood (or least squares) estimate, $\bar{\theta} = \bar{\beta}_1/\bar{\beta}_2$, is biased: $E(\hat{\theta}) - \theta \neq 0$. Taylor series approximations to the bias and variance of θ are given by the so-called delta method (Hinkley, 1977. But note that the formulae for the delta estimates of bias and variance that are given therein are incorrect. See Kotz and Johnson, 1982 under Delta Methods.) as:

$$E(\hat{\theta}) - \theta \approx \beta_1 Var(\bar{\beta}_2)/\beta_2^3 - Cov(\hat{\beta}_1, \bar{\beta}_2)/\beta_2^3$$

$$Var(\hat{\theta}) \approx Var(\bar{\beta}_1)/\beta_2^2 + \beta_1^2 Var(\hat{\beta}_2)/\beta_2^4 - 2\beta_1 Cov(\hat{\beta}_1, \hat{\beta}_2)/\beta_2^3.$$

Note that the bias, as well as the variance of $\hat{\theta}$, may be unacceptably large if the parameter estimates $\hat{\beta}_1$ and $\bar{\beta}_2$ are not well-determined by the sample data. This result has been generally ignored in current radiobiology praxis. In fact, the precision of the parameter estimates $\bar{\beta}_j$, as measured say, by $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$, $j = 1,2$, is not usually presented - or if presented, is often not correctly estimated - in most published reports. Of course, if the precision of the parameter estimates is high, the bias in $\hat{\theta}$ is negligible. However, in the reports that we assessed, the precision of the sample estimates of the model parameters was often quite low and the consequent bias was therefore, quite large, although this weakness in the sample estimate of θ was not remarked in the report itself.

A point estimate of θ in which the bias is less than that of $\hat{\theta}$, together with an interval estimate, i.e., $Var(\hat{\theta})$, can be obtained by the so-called weighted jackknife methods of Hinkley (1977). These methods have been shown to be especially useful when the data are unbalanced, i.e., the distribution of hat matrix diagonals for the model of the data at issue is markedly non-uniform. Hinkley (1977) has also shown that these methods can be used to obtain robust estimates of $\underline{\beta}$ and θ.

The weighted jackknife estimator of θ is constructed from the n row-deleted estimates of $\underline{\beta}$; that is, the set of $\underline{\bar{\beta}}_{(i)}$, $1 \leq i \leq n$. The i^th pseudovalue is defined to be $Q_i = \hat{\theta} - n(1-h_i)[\hat{\theta}_{(i)} - \hat{\theta}]$, $1 \leq i \leq n$. The jackknife estimator of θ is the mean of the n pseudovalues, $\hat{\theta}_J = n^{-1}\Sigma Q_i = \bar{\theta} - \Sigma(1-h_i)(\theta_{(i)} - \hat{\theta})$. Note that the weights are functions of the respective hat matrix diagonals, h_i, of the regression model.

The weighted jackknife estimate of the variance of $\hat{\theta}_J$ is

$$Var(\bar{\theta}_J) = n^{-1}(n-k)^{-1}\Sigma[Q_i - \hat{\beta}_j]^2.$$

Weber and Welsh (1983) remark that when the sample sizes, n, are small both the delta and weighted jackknife measures may "... grossly over-estimate the variance of $\hat{\theta}$ ..." The robust regression methods of Hinkley (1977) differ from those proposed by several others (e.g., Huber), "... in that 'harmless' large residuals are ignored: Q_i will not be extreme if $[h_i]$ is small." (Hinkley, 1977). See also Annex III.

For generalized linear models, $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, where the response $y_i = \mu_i + \varepsilon_i$, $1 \leq i \leq n$, has a (conditional) Normal distribution, and $\underline{\beta}$ is obtained by LS methods, the row-deleted estimates are

$$\underline{\hat{\beta}}_{(i)} = \underline{\hat{\beta}} - (X^T X)^{-1} e_i \underline{x}_i/(1-h_i),\ 1 \leq i \leq n.$$

For generalized linear models, in which the response has a Binomial or Poisson distribution and $\underline{\hat{\beta}}$ is obtained by IRLS methods the estimates, $\underline{\hat{\beta}}_{(i)}$, may be obtained by the same procedure from the transformed model $P\underline{y} = PX\underline{\beta} + P\underline{\varepsilon}$ where $P^T P = V^{-1}$ the (nxn) diagonal weight matrix of the observations estimated at the final iteration.

An alternative procedure for obtaining the row-deleted estimates, $\underline{\hat{\beta}}_{(i)}$, for generalized linear models of Poisson and Binomial responses is based on the mean shift outlier model described in section 7.1 above.

It is useful to note that the minimum expected loss (MELO) estimator of described by Zellner (1984) gives a shrunken estimate of θ that is quite similar in value to the reduced bias estimates obtained by the delta and weighted jackknife methods. The MELO estimator, $\bar{\theta}^*$, is derived from a Bayesian perspective. The MELO estimator is obtained as

$$\hat{\theta}^* = \hat{\theta}[1 + Cov(\hat{\beta}_1, \bar{\beta}_2)/\hat{\beta}_1\hat{\beta}_2]/[1 + Var(\hat{\beta}_2)/\hat{\beta}_2^2].$$

Zellner (1984) also gives a MELO estimator for the reciprocal of a regression coefficient, $\theta = 1/\beta_1$. See Annex II, part 5 and Annex III, part 4.

7.3.2 Direct estimates of $\theta = (\beta_1/\beta_2)$?

An alternative parameterization of the dose-response model of the LQ hypothesis from which the α/β ratio can be directly estimated has been proposed by several investigators. We illustrate it for cell survival models. Instead of the linear predictor, $\eta_i = \hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{\beta}_2 D_i^2$, the non-linear predictor, $\eta_i^* = \hat{\beta}_0 + \bar{\beta}_2(\gamma D_i + D_i^2)$, where $\gamma = \hat{\beta}_1/\hat{\beta}_2$ is an estimate of α/β, has been proposed. Whereas for the model with linear predictor η_i the so-called α/β ratio is a non-linear function $\theta = \beta_2/\beta_2$ of the parameters of the linear model with biased estimate $\hat{\theta} = \hat{\beta}_1/\bar{\beta}_2$ and in which the bias depends strongly on $\mathrm{Var}(\hat{\beta}_2)$, in the alternative parameterization, biased point and interval estimates of γ are obtained directly from the non-linear model.

The question is which is the more useful parameterization of the LQ model. Note that the two alternatives are reparameterizations of each other, "... that is, the parameters of one of them may be expressed as a function only of the parameters of the other model, without the expression containing the explanatory variables, the response variables, or the error term" (Ratkowsky, 1990). Ratkowsky (1983) has shown that the bias in the sample estimates of the parameters of the reparameterization is a strong function of the bias and variance of the rival parameterization. In the present case, the bias in γ is a function of Bias $(\underline{\hat{\beta}})$ and $\mathrm{Var}(\underline{\hat{\beta}})$. It is the case, of course, that the bias in $\hat{\theta}$ is also a function of $\mathrm{Var}(\underline{\hat{\beta}})$, as remarked above. It would seem to be incumbent on those who have proposed the rival parameterization to show that the bias (and variance) of γ are less than those of $\hat{\theta}_J(\equiv \hat{\beta}_1/\bar{\beta}_2)$. In fact, Ratkowsky (1983) has shown that the bias in the estimates of the parameters of some non-linear models may exceed 100% of the estimate! Moreover, the received method for obtaining reduced-bias estimators of the parameter vector of non-linear regression models is the Quenouille version of the jackknife. See Bard, "Non-linear Parameter Estimation", 1974. See Also Hinkley, 1977.

It is, of course, easy to show that the alternative, non-linear, parameterization of the LQ hypothesis $\eta_i = \beta_0 + \beta_2(\gamma D_i + D_i^2)$, is more non-linear than is the more familiar parameterization which is linear in the parameters: If the estimates of the respective parameter vectors of the two rivals are obtained by the same iterative least squares algorithm using the same initial estimates of the respective parameter vectors, the model with the non-linear predictor requires several-fold more iterations to achieve parameter estimates of the same stability; the difference in the number of iterations is a sample measure of the difference in degrees of non-linearity of the respective parameterizations. (N.B. It can be shown that for a linear model, the usual Gauss-Newton iterative least-squares methods that are required to estimate the parameters of non-linear models, will converge to the final estimates of the model parameters in a single step - from any set of initial estimates (Ratkowsky, 1983). Moreover, these estimates will be unbiased, minimum variance estimators.)

Note also that, as we have demonstrated, if the linear parameterization, η_i, "fits" and $\underline{\beta}$ is precisely estimated, then the bias present in the non-linear function $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$ will be small. But, whatever its size, the bias in $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$ is readily estimable from the estimate $\underline{\hat{\beta}}$ and the diagnostics, h_i and $\underline{\hat{\beta}}_{(i)}$, that are, or should be, routinely constructed in the analysis of any regression model. Moreover, goodness-of-fit measures including the more sophisticated regression diagnostics such as the hat-matrix diagonals h_i and the row-deleted estimates $\underline{\hat{\beta}}_{(i)}$, are much more readily estimated - and interpreted - for linear, or generalized linear, models than for non-linear models (Welsch, 1986). Indeed, the serious work in developing algorithms for computing regression diagnostics for non-linear models of non-Normally distributed data has barely begun. Note as well that the estimation of the bias in ML estimates of the parameters in non-linear regression models is rather more complex than that for $\hat{\theta}$ $(=\hat{\beta}_1/\bar{\beta}_2)$. See Box, M.J. (1971).

However, there is another still more consequential issue. The alternative non-linear parameterization must be based on the assumption that the LQ model is appropriate for the data at hand. (As Ratkowsky, 1983 has remarked anent the deployment of non-linear models in small data sets: "The usual modelling situation is that a model is adopted because some theory and/or empirical evidence from the use of that model over many data sets indicates that model is appropriate.") But in the case of the LQ model neither the theory nor the empirical evidence for it are compelling. Therefore, the concordance of the model must be examined in each deployment.

And, goodness-of-fit is best checked in the linear parameterization. It must not be assumed a priori that the LQ model is appropriate in every study in which it is deployed. As we have shown elsewhere (See Annexes II and III) the LQ model may either overfit a set of data (as is the case for the BEIR III leukemia incidence data) or it may underfit a set of data (as is the case for the rat bone marrow stem cell survival data). It should be recalled that, "In fact, most polynomial models, whether linear, quadratic, or higher order, should be viewed as being in the nature of a low-order power series expansion of the response function about the centroid of the experimental region." D. Marquardt, 1980.

Since the analysis of the linear parameterization provides the hat-matrix diagonals, h_i, and row-deleted estimates, $\hat{\beta}_{(i)}$, $1 \leq i \leq n$, from which the jackknife estimates, $\hat{\theta}_J$, and $Var(\hat{\theta}_J)$ can be readily constructed there would seem to be no reason to re-fit the data using non-linear parameterization in order to obtain "direct estimates" of γ - especially since these may well be strongly biased themselves. (N.B. Cox, 1990 has recently described a method for obtaining interval estimates of the ratio, $\theta = \beta_1/\beta_2$, that is an alternative to the delta method, Fieller's theorem and Hinkley's weighted jackknife. It is based on a re-parameterization of the model - a "non-linearization" of a linear model. For example, the Poisson linear LQ model of mutagenesis would be re-written as follows:

$$m_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 = \beta_0 + \beta_2(\gamma D_i + D_i^2)$$

Then maximum likelihood estimates of θ are obtained directly from the non-linear model. It would seem that this procedure suffers from the weaknesses described above.)

This brief review of the current concepts, methods, and criteria of statistical modelling provides the required perspective from which to assess the received praxis in radiation biology. For a more detailed account the reader is referred to Annexes I-IV and the references cited therein. It should be noted that the software (save that required to implement the SIMPLEX procedure and the non-linear Poisson regression model) that was necessary to implement the statistical methods and criteria described above and that were deployed in the secondary analyses reported in Annexes I-IV was written by the principal author in the BASIC language.

7.4 Roles for the non-linear function $\theta = (\beta_1/\beta_2)$

The solutions to several consequential problems in radiation biology, radiation epidemiology, and radiation oncology devolve into inferences on the ratio, say $\theta = \beta_j/\beta_k$, of two regression coefficients, say β_j and β_k. For example, the so-called cross-over dose is the ratio, $\theta_1 = \beta_1/\beta_2$, of the coefficient of the linear term in dose, D, to that of the quadratic term D^2 in the so-called LQ-L model of radiation leukemogenesis. The dose-rate effectiveness factor, DREF, is the ratio, $\theta_2 = \beta_1(L)/\beta_1(LQ)$ of the coefficient of the linear term in the linear model of leukemogenesis to the coefficient of the cognate term in the rival linear-quadratic (LQ) model of the same set of responses. The so-called neutron RBE for induction of mammary neoplasia in rats can be shown (See Annex III, part 6) to be $10^{\theta}3$ where $\theta_3 = -\beta_2/\beta_1$, β_2 is the coefficient of a (0,1) indicator variable, x_2, for LET (neutron/gamma), β_1 is the coefficient of the log radiation dose (either gamma or neutron), x_1, in the probit model $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where $z = \Phi^{-1}(\pi)$, $\Phi(\ )$ is the Normal distribution function and $0 \leq \pi \leq 1$ is the proportion of responders at (x_1, x_2) in the probit model of response in the pooled samples of gamma and neutron irradiated animals. The neutron RBE for human leukemogenesis is defined (BEIR III, 1980) to be the ratio $\theta_4 = \beta_1(L)/\beta_2(L)$ where $\beta_1(L)$ is the coefficient of the gamma dose D_γ in the L-L model of leukemia incidence and $\beta_2(L)$ is the coefficient of the neutron dose D_n in that model. A similar definition obtains for the neutron RBE for breast cancer in the females of the LSS sample. See BEIR III, 1980. The so-called α/β ratio of the linear-quadratic (LQ) model of clonogenic survival in cells exposed to single-fractions of low-LET radiation is the ratio $\theta_5 = \beta_1/\beta_2$ where β_1 is the coefficient of dose D and β_2 is the coefficient of D^2 in the LQ model.

Since the ratio $\theta_5 = \alpha/\beta$ is unique only to within a multiplicative constant it would appear to be a dubious practice to invest it with too much discriminatory significance. Nonetheless, the α/β ratio is currently often deployed to discriminate between early and late functional responses in normal tissues/organs: $\alpha/\beta = 3$ (late); $\alpha/\beta = 10$ (early, including tumor response). (Fowler 1984a, b) Such deployment is based on the (questionable) assumption that an observed level of a binary effect is uniquely determined by an unobserved level of survival in an unobservable (and indefinable) cell population. Empirical evidence against this practice can be found in the example of two observable cell populations for which the α/β ratios differ by only about 10% but have markedly different survival curves - the respective survivals at 1.5 Gy differ by a factor of 2.0 - that is presented in Fig. 27c of section 7.10 of this report.

However, as well as the inferential errors which may encumber the practical deployments of these ratios, there are important aspects of the sample estimates of these ratios themselves that are of concern, namely, the bias, the variance and the lability of the estimate. See Annexes II, part 5 and III, part 4. We will discuss later the issue of the lability of these estimates; that is, the sensitivity of the point estimate $\hat{\theta}$ to the deletion of a single observation from the sample. See Annex II, part 5, Annex III, part 4 and Figs. 23c and 25a-25c of this report. The issues of bias and variance of the sample estimates $\hat{\theta}$ of these ratios will be briefly summarized here.

Although the variance of - or confidence limits on - the estimate of one of these ratios may sometimes accompany the point estimate in a published report more often than not it is omitted. And we have found at least one published report, NCRP 64, in which the variance of the estimate of $\theta = \alpha/\beta$ is incorrectly estimated (the term in $\mathrm{Cov}(\alpha, \beta)$ is omitted from the delta estimate of $\mathrm{Var}(\theta)$. See Annex III, part 5.)

However, there have been no estimates of the bias in the point estimates of these ratios ever published. Indeed, the question of the bias of the estimate has, apparently, not been considered hitherto. Nonetheless, as remarked above, because $\theta = f(\underline{\beta}) = \beta_j/\beta_k$, say, is a non-linear function of the parameters of a regression model, the maximum likelihood (or least squares) estimate $\hat{\theta} = \hat{\beta}_j/\hat{\beta}_k$ is biased. And the size of the bias depends strongly on the variance of the denominator. As also discussed above, there are three methods for obtaining reduced bias estimates of $\theta = \beta_j/\beta_k$: the delta method, the weighted jackknife method and the minimum expected loss (MELO) method. The first two methods also yield estimates of the variance, $\mathrm{Var}(\hat{\beta}_j/\hat{\beta}_k)$.

As shown in the two Annexes cited above, the bias and variance of the sample estimates of these ratios can be quite large. For example the sample estimate of the so-called cross-over dose for leukemia incidence is given in the 1985 Report of the Working Group of the NIH as $\theta = 117$ rad. However, the reduced bias estimates of θ obtained by the delta, weighted jackknife and MELO methods are, respectively, 3 rad(!), 18 rad and 29 rad. The 0.68 confidence limits on θ as given by the delta, Fieller's theorem, and weighted jackknife methods are (-91, 271), (3.89, 647) and (-196, 232), respectively. The bias and variance of the estimate $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2 = 117$ rad are so large because neither the numerator, $\hat{\beta}_1$, nor denominator, $\hat{\beta}_2$, exceed their respective standard errors by factors greater than 1.50. (Indeed, $\bar{\beta}_1/\surd\mathrm{Var}(\hat{\beta}_1) \sim 1.0$.) In the LQ-L model of leukemia incidence, the covariance matrix, $\mathrm{Var}(\underline{\hat{\beta}})$, is inflated because the model is over-fit - the term in D_γ^2 is not needed to achieve a statistically adequate fit to the data. (Indeed, it can be shown that all three models listed in Tables V-8 and V-9 of the BEIR III (1980) Report, overfit the leukemia incidence data since each of them includes an indicator variable C = 0,1 that identifies the city, Hiroshima and Nagasaki. But it can be shown that the estimate of the coefficient of C is less than its standard error for each of the three rivals: LQ-L, L-L, and Q-L. See Herbert, 1986b.) This issue is discussed at some length in Annex III, part 4 of this report. However, it is evident from Table V-8 of the BEIR III report itself that the LQ-L model overfits the LSS leukemia incidence data since the decrement in the chi-squared statistics for the LQ-L model and either of the rival models, L-L or Q-L, is obviously much less than the 0.95 quantile, $\chi^2 0.95(1) = 3.84$, of the chi-squared distribution for 1 degree of freedom (df).

Similarly, for the estimate of $\theta = \alpha/\beta$ for an LQ model of survival curve for rat stem cells in Annex II, part 5 we find that $\theta = 5.88$ and the reduced bias estimates are 5.34, 3.25 and 5.11 as estimated by the delta, weighted jackknife and MELO methods, respectively. It should be recalled that the weighted jackknife estimates are to be preferred over the delta and MELO estimates since the latter require the assumption that the parameter estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$, are equal to the respective parameters, β_1 and β_2.

We had remarked earlier that it appears that the stochastic part, ϵ_i, in the radiation response $y_i = \mu_i + \epsilon_i$ is regularly ignored. It thus comes as no surprise to find that the inherent uncertainty and ambiguity which must encumber any estimates and inferences constructed from noisy data - and biological data are very noisy - is also ignored. That is, the literature includes very few papers in which the uncertainties, Var($\hat{\underline{\beta}}$) and Var(f($\hat{\underline{\beta}}$)), in the sample estimates, $\hat{\underline{\beta}}$ and f($\hat{\underline{\beta}}$), of the parameter vector $\underline{\beta}$ and functions, f($\underline{\beta}$), of the parameter vector, such as the response (f($\underline{\beta}$) = $\underline{x}^T\underline{\beta}$), the α/β ratio (f($\underline{\beta}$) = β_1/β_2), etc., are reported.

Similarly, we find that there are no published papers which report the uncertainty, say ψ, in the a priori information on the parameter vector as described, say, by the constraint $\underline{r} = R\underline{\beta} + \underline{v}$, E($\underline{v}$) = 0, Var($\underline{v}$) = ψ. Such omissions suggest that either we do not know how well we know what we know, or that we believe it to be revealed truth. Either circumstance will, of course, "... wonderfully obstruct the understanding."

Of course, it is commonly the case that the prior information on $\underline{\beta}$ that is described by the matrices $\underline{r}$ and R is more readily specified than is the precision of that information, described by the matrix ψ. Thus, it is good practice to obtain posterior estimates, $\hat{\underline{\beta}}^{**}$, for a range of values of ψ and then determine how sensitive these estimates are to changes in ψ, i.e., to perform a sensitivity test (Leamer, 1986). And, "... [We] recommend that one always perform sensitivity analyses. That is, one should determine if moderate changes in one's prior beliefs or one's analytic procedures ... would lead to large changes in one's inferences about the parameter of interest" (Robins and Greenland, 1986). However, such sensitivity analyses are never reported in those papers in which such prior information is deployed. It appears that the investigators invariably assume, for example, that ψ = [0], the null matrix. Figure 4 in Annex IV, part 3 describes the effect on the posterior estimates $\hat{\underline{\beta}}^{**}$ of varying the size of ψ in a sensitivity test of an isoeffect model. (See also Fig. 36b of this report.)

Some more detailed accounts, with illustrations, of the several weaknesses of the dose-response studies reported in the current literature that have been disclosed to our re-analyses seems in order now. These weaknesses are 1) the misspecification of either, or both, the deterministic and the random parts of the model (ontological weaknesses) and 2) the failure to systematically assess the consistency of the model with the data and with prior beliefs (epistemological weaknesses). Each of these weaknesses is discussed more fully, together with a concise exposition of the statistical methodology through which it was disclosed, in the Annexes II-IV of the present Report.

7.5 Models of dose-response and isoeffect

"... the ill and unfit choice of words, wonderfully obstructs the understanding."

F. Bacon, 1620

"Experienced scientists know that the mathematical instrument that is chosen in each case to express a given objective pattern depends not only on the sort of phenomena themselves that are concerned but also on the scientist's own mathematical ability and equipment."

M. Bunge, 1959

"Modellers should realize that it is ludicrous to attempt to fit a model that graphically describes a certain shape to data that does not conform to that shape."

D. Ratkowsky, 1990

"The size of the confidence limits is inversely proportional to the quality of the data used to make the estimate and directly proportional to the amount of extrapolation involved. This important information is lost if the confidence limits and best estimates are not routinely reported. The width of the confidence interval is one of the best measures risk assessors, and risk managers have to evaluate the quality of the estimates of potential risks. It is important to distinguish between those situations in which the risk is precisely estimated and those in which it is not."

C. Park and R. Snee, 1983

"It is now generally recognized that we can have a reasonably acceptable explanation of a phenomenon in the absence of an ability to predict its occurrence and that we may be able to predict phenomena in the absence of a fully satisfactory explanation."

M. Oaks, 1990

Although the fundamental professional problem of clinical radiation therapy is to *maximize the probability of uncomplicated control of disease*, the received radiobiological models cannot well serve to inform solutions to that problem; indeed, they cannot serve even to well illustrate the solution. This is because the fundamental problem requires estimates of the conditional *joint probability* of the concomitant occurrence, in the target volume, of at least two quantal responses: *ablation* of tumor and complication of normal tissues. This may be usefully expressed as $P(S|\underline{x}^T)$ where S is the joint event, $S = E_1$ and $\bar{E}_2$, treatment success, E_1 is the binary event, ablation of tumor and E_2 is the (binary) event, complication of normal tissue, ($\bar{E}_j$ is the complementary event to E_j). Estimates of $P(S|\underline{x}^T)$ must be obtained from models of *dose-response*. However, all previous and current efforts in radiobiological modelling have been directed to the construction of models of *isoeffect*. But there is, as Casti (1989) has remarked, a "... deep epistemological issue in the theory of models, namely, the distinction between *causality* and *determinism*." For example, "... the ideal gas law asserts that once we know any two of the three observables P, V, and T, the remaining observable is *determined* by the other two, but not *caused* by them." This is one crucial distinction between the dose-response models, which are *causal*, and isoeffect models, which are *deterministic*, i.e., once we know N and T the isoeffect dose, $D(\pi)$, is *determined* - but not caused - by the levels of N and T, whereas, a given level of a response, say π, $0 \leq \pi \leq 1$, for a given tissue, is *caused* by a given set of levels of D, N, and T. But, as Oakes (1990) has remarked, "... in the development of a science it is *causal relationships* which are at a premium." (emphasis added)

It is, of course, the case that while models of isoeffect can be derived from models of dose-response, the converse does *not* obtain. In fact, the dose-response model can generate *families* of isoeffect models, one for each level of the conditional probability, $\pi = P(E_j|\underline{x}^T)$. See Appendix II. (It is, of course, also the case that both epidemiological and laboratory studies of the so-called low-dose effects in biological systems, mutagenesis and carcinogenesis, rely on causal *dose-response* models, while clinical and laboratory studies of high dose effects such as toxicity, rely on deterministic *isoeffect* models. But in both studies, the matter at issue is to obtain "useful" estimates of the conditional probability, $P(E|D)$ of the occurrence of the binary event E at dose D; i.e., a causal model is required.

Note however that in the *received literature* (apparently), on the authority of common practice, the *dose-response surface* for the events E_1 - or E_2 - that is *immanent* in the sample data on any radiation dose-response relationship in which the *stochastic* part, ϵ_i, of the response has a (non-Normal) Binomial distribution and the *deterministic* part, μ_i, is multivariate, is invariably collapsed - projected - into an *isoeffect curve*, before the deployment of any computational procedures to obtain estimates and inferences. This transmogrifies the primary problem to be addressed by models of radiation therapy, namely that of accounting, in terms of a multivariate model, for the observed variation in the conditional *probability of occurrence of a quantal response*, say the proportion π, with dose, say D, *and other covariates*, into the *inverse* - and derivative - problem of accounting for the variation of the dose, $D(\pi)$, that elicits a specified level of response, π, with other covariates. We have referred to this as the "isoeffect transformation" since it serves in the same function for quantal data in which the response has a Binomial distribution, as does the "survival transformation" for count data in which the response has a Poisson distribution. That is, it serves to transform the metric of the response variable so that ordinary least squares methods can be deployed for estimation of the parameter vectors of the received models.

However, for data in which the observed responses have either Binomial or Poisson distributions, the inherent metrical nature (proportions or counts, respectively) of the response itself *prescribes* the transformation of response that is required for the appropriate generalized linear model (Dobson, 1983; McCullagh and Nelder, 1989; Maindonald, 1984) of *dose-response*. The

Binomial and Poisson generalized linear models are described below in sections 7.5.1 and 7.5.2, respectively. It must be emphasized that these are not the models usually reported in the radiobiology literature. It is necessary that these dose-response equations be included here in order to better explicate, subsequently, the ways in which the current modelling praxis of the radiobiology peer-group differs from that of proper statistical modelling.

It is, of course, the case, that for a set of clinical data the level of the parameter, π, of one of the Binomial responses of interest, say complication of normal tissue, may not vary much (by intent of the attending physician which, in turn, is determined by "the standard of practice in the community") about a fixed level, say $\pi = \pi^*$, over the sample of observations. π^* represents, "... the amount of radiation damage that is acceptable for purposes of curing a cancer," and "... remains one of personal philosophy." according to Rubin and Casarett (1973). Therefore, it is inherently difficult to construct from such data, by any set of methods and criteria, useful models that can account for the observed conditional variation in the proportion, π, of responders in either normal or tumor tissues with dose, etc. (dose-response models), over the full range, $0 \leq \pi \leq 1$, such as described in section 7.5.1. See for example, Appendix II, Fig. 2. (One remedy for this inherent weakness in clinical dose-response data is through the deployment of Bayesian hierarchical meta-analytic methods by means of which clinical studies can "borrow strength" from parallel animal studies and is discussed below in part 14, "What shall we do now?".) However, models constructed on covariates such as fractions (N) and/or time (T), of the variation in the dose $D(\pi^*)$ that is required to elicit a specified level, π^*, of response (isoeffect models) can be - and, of course, regularly are - readily constructed from clinical data. See Annex IV, Part 3 and Appendix II.

Let us expand a bit on the notion that dose-response models are causal. Thus, for the Power-law hypothesis we have the dose-response model, say $z_i = \beta_0 + \beta_1 \log D_i + \beta_2 \log N_i + \beta_3 \log T_i$, where z_i is the probit (or logit) transform of the proportion, π_i, of responders for n_i at risk at (D_i, N_i, T_i), that is, the level of response is caused - in part - by the levels of the treatment variables. However, the cognate isoeffect model is deterministic but not causal: $\log D_i(\pi^*) = \alpha_0 + \alpha_1 \log N_i + \alpha_2 \log T_i$. The level of dose, $D(\pi^*)$, that elicits the proportion, π^*, of response is determined, but not caused, by the levels of N and T. Some of the problems we have described in current clinical radiobiological praxis may be due to the failure of many investigators to clearly distinguish between causal and deterministic models. For example, it is, unfortunately, also the case that this practice of constructing models of isoeffect rather than models of dose-response has led to the design and analysis of animal experiments in support of only isoeffect models even though in these experimental observations the parameter π of the Binomial distribution of observed response mkay, and often does, vary over a sufficient range of response, say, $0 \leq \pi \leq 1.0$, so that a useful dose-response model could be constructed directly from the data. In that case, the inverse isoeffect model could be derived subsequently as - or if - desired. Therefore, in the received practice of radiobiological modelling, much of the empirical information on dose-response that is obtained (at no little effort and expense) by experiment is often quite ruthlessly discarded in order to summarize the sample data by the cognate isoeffect curve. The dose-response surface is, effectively, projected, or collapsed, into the isoeffect curve - as, for example, in the F_e-plots for the LQ model. These latter are often constructed from experimental data in which the observed conditional probability of the Binomial response varies over the full range, $0 \leq \pi \leq 1.0$, in each of a set of $n > 1$ dose-response experiments. However, the derived F_e-plot captures only the response at a single level, say $\pi^* = 0.5$ in each of the experiments. An LQ regression model of isoeffect is subsequently constructed by Least Squares methods from these constricted "data". See, for example, Figs. 9, 10, and Annex II, part 3 (in which n=7).

Let us examine Fig. 9 and 10 more closely. The set of experimental data that were represented in Fig. 5 of Tucker and Thames 1983 (our Fig. 9) appeared to us to be rather odd. For example, there is a marked lack of uniformity in the number of dose-groups assigned to each of the 7 levels of fractions, N: At N = 1, 3,and 15 there are 3 dose-groups, at N = 2, there are 4, at N = 5, 10, and 20 there are 2. There is also an excessive amount of the data that reside in the extreme responses: At 5 observations we have $r_i/n_i = 1$; at 7 observations, $r_i/n_i = 0$; at only

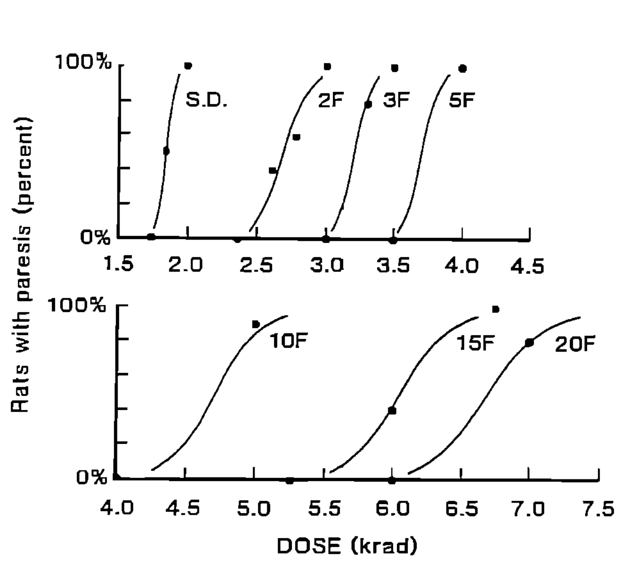


Fig 9. Plot of a quantal dose-response experiment. The end-point is hind-leg paresis in rats. The experiment is typical of its class. The figure is reproduced (with permission) from Tucker and Thames (1983) who cite the van der Kogel (1979) experiment as the source of these data. The specification of the numbers at risk, n_i, at the ith treatment regimen (D_i, N_i, T_i), $1 \leq i \leq 19$, was infered from the somewhat ambiguous statement in Tucker and Thames (1983): "Assuming that the average number of doses used to obtain each quantal curve was $n_D = 3$ and that $n_A = 10$ animals were irradiated per dose ..." We have assumed in our analyses that $n_i = 10$. Several features of this experiment must immediately strike the reader as exceedingly odd:

a) An excess, 12 of 19 or 63%, of the data resides in the extreme responses, $r_i = 0$ or $r_i = n_i$, where r_i is the number of responders in n_i at risk at $\underline{x}_i^T$.

b) In 9 of the 19 regimens, the dose per fraction, $d = \bar{D}/N$, exceeds 10 Gy, the stipulated upper limit of validity of the multifraction ($N \geq 1$) LQ model. (Fowler, 1984).

c) In 19 of the 19 regimens the dose per fraction, $d = D/N$, lies <u>outside</u> the range of doses per fraction in the radiotherapy range.

d) It can readily be shown that the covariates N and T are highly correlated, i.e., the data are <u>multicollinear</u>.

e) The number of levels of dose, D, at each of the seven levels of N varies between two and three save for the experiment at N=2 for which there are four levels of dose. The average number of levels of dose is $n_D = 2.71$, whereas to adequately determine the location and shape of a binary dose-response curve, $n_D = 5$ levels are required. The experimental design is obviously flawed.

The remarks of R. A. Fisher seem appropriate: "If the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too."

We later learned (van der Kogel, 1987) that for the experiment that was actually performed, the numbers of animals at risk differed substantially from that reported by Tucker and Thames (1983): $n_D = 5$ and $n_A = 5$. Thus, the data in the Tucker and Thames (1983) report must be regarded as <u>non-experiential</u>: "Something ... admitted or assumed for specific purposes." ("datum. 1a: Something that is given either from being experientially encountered or from being admitted or assumed for specific purposes: a fact or principle granted or presented: something upon which an inference or an argument is based or from which an intellectual system of any sort is constructed." Webster's Third New International Dictionary, 1966). But even $n_i = 10$ animals at risk at the ith level of treatment is too small - by a factor of 3-5. We note that this is also the case for most Phase III clinical trials: $n_i = 20$ where $n_i = 100$-200 is required. (See Zelen, 1982.) This is an instance of deployment of the principle of parsimony in making observations.

The experiment described by Tucker and Thames (1983) must be regarded as a "thought-experiment" since the data are "non-experiential".

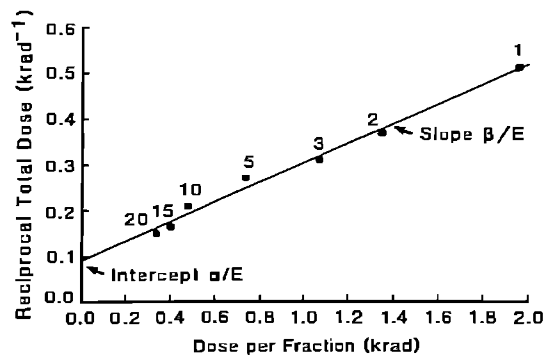


Fig. 10a. F_e-plot constructed from the data of Fig. 9. The points are ED50s - they give the estimated levels of $(D_i^{-1}, D_i/N_i)$ required to elicit the response - hind-leg paresis - in 100 π = 50% of the animals so exposed. Thus, the line represents the $\pi = 0.50$ isoeffect curve. The F_e-plot maps the dose-response curves of Fig. 9 in which $0 \leq \pi_i \leq 1$, $1 \leq i \leq 7$, into the isoeffect curve, $\pi_i = 0.50$, $1 \leq i \leq 7$, via an "isoeffect transformation". The Tucker and Thames (1983) point estimate of the α/β ratio obtained from the F_e-plot of Fig. 10 by fitting these "data" to the equation, $D^{-1}(\pi) = (\alpha/E) + (\beta/E)D(\pi)/N$, where $E = -\ln S$, by ordinary least squares methods, is $(\alpha/E)/(\beta/E) = 0.42$. Although not provided in Tucker and Thames (1983) we have estimated (by Fieller's theorem) the 0.95 confidence limits on this estimate of α/β to be (0.269, 0.609).

It is remarked in Tucker and Thames (1983) that their estimates of the respective slopes of each of the dose-response curves of (our) Fig. 9 and thus the estimates of ED50 in Fig. 10 were obtained by fitting the data "by eye" (a curious practice that yields neither estimates of goodness-of-fit of the model nor of the precision of the parameter estimates).

Note further that the dose-response data of Fig. 9 have been projected - collapsed - into the isoeffect "data" of Fig. 10, a projection that, although it has the authority of common practice, nonetheless, obviously discards important information on dose-response. Apparently, the "isoeffect transformation" such as described in Fig. 10a is commonly deployed to enable the investigator to construct regression models of Binomial dose response data by ordinary least squares methods instead of the iterative reweighted least squares methods required by the logit or probit transformations of Binomial data. We shall see that it is a poor choice of transformation.

Nonetheless, this method of estimation of α/β is currently recommended: "Although the reciprocal total dose or F_e-plot (Douglas and Fowler, 1976) is not the most accurate way to calculate α/β, it is the easiest method and gives a fairly accurate value if the data are good, but only an optimistic estimate of the error range." (Fowler, 1989). However, the statistically appropriate (generalized linear) model of the multifraction LQ hypothesis of the dose-response data of Fig. 9 is obtained from the probit transformation, $z_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2/N_i$, where $z_i = \Phi^{-1}(\pi_i)$ is the probit transform (see Fig. 11a). This yields the point estimate $\hat{\alpha}/\beta = \bar{\beta}_1/\bar{\beta}_2 = 0.328$. The sample estimates of $\underline{\beta}$ for the probit model are obtained by iteratively re-weighted least squares (IRLS) methods.

N.B.: It can be shown that, on the normative statistical criteria of "fit", that derive from comparisons of the response observed, r_i of n_i, with that predicted by the model, $n_i\hat{\pi}_i$ (at $\underline{x}_i^T$), the (statistically appropriate) probit model (M2) of the multifraction LQ hypothesis does not fit the data of Fig. 9. See text and also Annex II, part 3.

N.B.: Since S is unknown - neither the level of survival nor the cell population to which it refers are known - the LQ model deployed in the F_e-plot is rather devoid of empirical content.

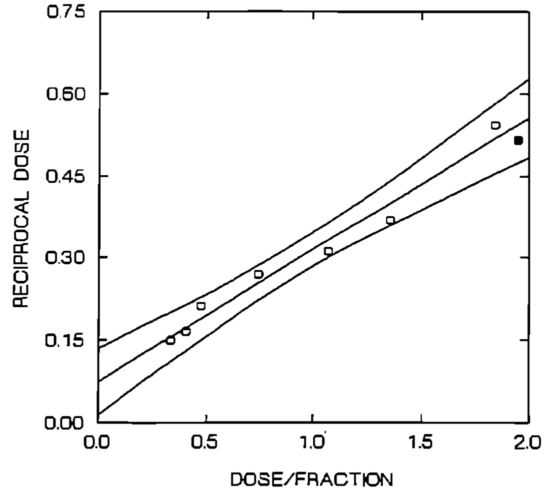


Fig. 10b is an F_e-plot taken from the data of Fig. 9. The data differ from the cognate plot of Fig. 10a in that in the latter there is an error in the location of the point $(D_1^{-1}, D_1/N_1)$. The erroneous point is identified by the filled symbol. The plot differs further from that of Fig. 10a in that in Fig. 10b the isoeffect curve is appropriately circumscribed by the appropriate 0.95 CL.

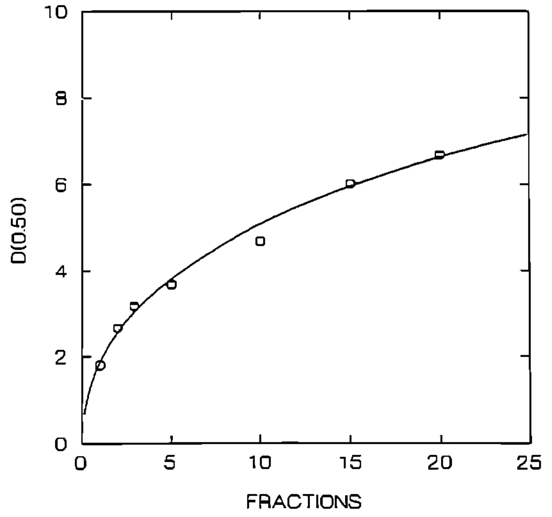


Fig. 10c is a plot of the correct $\pi = 0.50$ isoeffect curve for the LQ hypothesis constructed from the dose-response data of Fig. 9. The generalized linear dose-response model of this hypothesis is $z = \beta_0 + \beta_1 D + \beta_2 D^2/N$ where z is the probit (or logit) transform of a Binomial response. The (reparameterized) cognate isoeffect model derived therefrom is $D_i(0.50) = [-\alpha_1 + \sqrt{}(\alpha_1^2 - 4\alpha_0/N_i)] /(2/N_i)$, $1 \leq i \leq n = 7$. The parameter estimates (standard errors) are $\bar{\alpha}_0 = 3.866(0.482)$, $\bar{\alpha}_1 = 0.253(0.086)$. For this model $\bar{\alpha}_1$ is the point estimate of α/β with 0.95 CL (0.032, 0.474).

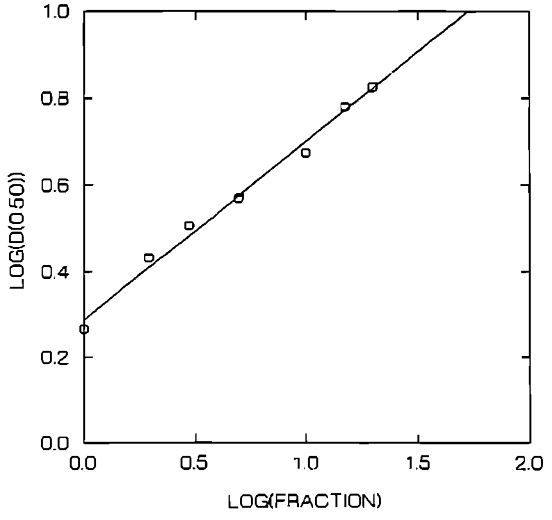


Fig. 10d is a plot of the correct $\pi = 0.50$ isoeffect curve for the Power-law hypothesis constructed on the isoeffect data $(D_i(0.50), N_i)$, $1 \leq i \leq n = 7$, of Fig. 9. The generalized linear dose-response model of the Power-law hypothesis is $z = \beta_0 + \beta_1 \log D + \beta_2 \log N + \beta_3 \log T$, where z is the probit (or logit) transform of a Binomial response. The (reparameterized) cognate isoeffect model is $\log D_i(0.50) = \alpha_0 + \alpha_1 \log N_i$, $1 \leq i \leq n = 7$. The parameter estimate with (standard errors) are $\alpha_0 = 0.288$ (0.015) and $\alpha_1 = 0.413$ (0.018). This gives the model $D(0.50) = 1.925N^{0.413}$ krad.

7 observations do we have $0 \leq r_i/n_i < 1$; $1 \leq i \leq 19$. Here r_i is the number of responders in n_i animals at risk. (The presence of so many extreme responses suggests that the n_i are too small. In a properly designed experiment, $30 \leq n_i \leq 50$, $1 \leq i \leq n$. See Finney, 1971b.) And so forth. For these reasons, we sought and obtained a copy of the original data from van der Kogel (private communication). We found that the experiment described in Fig. 5 of Tucker and Thames (our Fig. 9) is quite different from the experiment actually performed by van der Kogel. For example, there were actually 4-5 dose-groups assigned to each of the 7 levels of N rather than the numbers given above. Moreover, there were no more than $n_i = 6$ (and most often only $n_i = 5$) animals at risk in each dose-group rather than the $n_i = 10$ described in Travis and Tucker, 1983. Thus, the data in Travis and Tucker appear to be "non-experiential". Their Fig. 5 (our Fig. 9) describes a "thought experiment" (See Annex II, Part 4. "Some Thoughts on a 'Thought Experiment'." We also obtained from van der Kogel the number of days of treatment, T_i, as well as the number of fractions, N_i. In this manner, we acquired a complete set of observations: r_i, n_i; D_i, N_i, T_i, $1 \leq i \leq n = 32$.

However, the purposes of our secondary analyses required that we use the Tucker and Thames version of the van der Kogel data, save that we included the values of T_i, $1 \leq i \leq 19$, corresponding to the levels of D_i and N_i of Fig. 9. And, of course, we used $n_i = 10$, $1 \leq i \leq 19$, as described by Tucker and Thames. Thus, we were able to construct dose-response models of the two rival hypotheses, LQ and Power-law, on the data described in Fig. 5 of Tucker and Thames (our Fig. 9). The construction, criticism, and comparison of these models is described below in section 7.9 and Fig. 14-Fig. 19. For the present we examine several other aspects of these models of these data.

It must be remarked again, as well, that Fig. 2 of Tucker and Thames (our Fig. 10a) and indeed, all other F_e-plots also strike us as a rather odd sort of regression equation since the dose variable appears as both a dependent and an independent variable: $D^{-1} = f(D)$. (Still "odder", perhaps, is the practice by which the 19 <u>observations</u> on dose-response in Fig. 9 are reduced to 7 <u>estimates</u> of isoeffect in Fig. 10a <u>before</u> any calculations are made.) For those investigators who use the LQ hypothesis to direct their studies and interpret their findings, the usual purpose of constructing the F_e-plots is, apparently, to obtain, from isoeffect models, sample estimates of the ratio α/β for the LQ dose-response model. The estimates of α/β are used, for example, to identify a tissue as the possible site of <u>early</u> (α/β ~ 10 Gy) or <u>late</u> (α/β ~ 3 Gy) radiation effects (Fowler, 1989). In sections 7.3, 7.4, and 7.9 and in Annex II, parts 3 and 5, and Annex III, part 3, of the present report, we consider at some length the problem of obtaining appropriate estimates of the <u>bias</u> and <u>variance</u> of the sample estimates of α/β ratios obtained from <u>dose-response</u> models of LQ hypotheses (part of the problem here is that the ratio of two <u>estimates</u> of regression coefficients is a <u>biased estimate</u> of the ratio of the coefficients). It is important to note at once that we have shown that the <u>point estimates</u> of α/β obtained from the appropriate generalized linear model of Binomial dose-response data may differ by consequential amounts from the point estimate obtained from the cognate isoeffect F_e-plot.

For present purposes, however, we "go along" with this peculiar practice of the radiobiology peer group but offer, as well as a point estimate of α/β, a <u>set</u> estimate; in this case, an estimate of the 0.95 confidence limits on that ratio obtained from the F_e-plot of Fig. 2 of Tucker and Thames, 1983 - our Fig. 10a. (We note that interval estimates were omitted in the original report from which Figs. 9 and 10a were taken - as is often the case in the literature of the LQ model.) The estimates of the 0.95 confidence limits on sample estimate $\alpha/\beta = 0.421$ obtained from Fig. 10a are (0.253, 0.589) and (0.269, 0.609). These were obtained by the so-called delta method (Hinkley, 1977) and Fieller's theorem (Finney, 1971), respectively, as described above.

However, it must be recalled that Fowler has cautioned, concerning point estimates of α/β obtained from the F_e-plot, that, "There is a problem in assigning experimental error limits to the values of α/β ... In practice a visual indication of the spread of α/β values is, of course, provided by the deviation of points from the line drawn through them" (Fowler, 1984). More recently, this warning was repeated: "Although the reciprocal total dose or F_e-plot (Douglas & Fowler, 1976) is not the most accurate way to calculate α/β, it is the easiest method and gives a fairly accurate value if the data are good, but only an optimistic estimate of the error range" (Fowler, 1989). But in neither the 1984 nor the 1989 reviews by Fowler were any quantitative estimates of the "error range" given. Two insistent questions arise, of course, at once: a) How accurate is "fairly accurate"? and b) what, precisely, is meant by "optimistic"? For example, in Annex II, part 3 the point estimate of α/β obtained from the appropriate generalized linear model of the LQ hypothesis for the data of Fig. 9 is $\bar{\alpha}/\hat{\beta}$ = 0.328. This model is the multivariate probit model $z_i = \underline{x}_i{}^T\underline{\hat{\beta}}$, or, $z_i = \hat{\beta}_0 + \bar{\beta}_1 D_i + \hat{\beta}_2 D_i{}^2/N_i$, $1 \leq i \leq n$ (=19), where $z_i = \Phi^{-1}(\pi_i)$ and $\hat{\alpha}/\hat{\beta} = \hat{\beta}_1/\hat{\beta}_2$. But $\bar{\alpha}/\bar{\beta}$ = 0.421 is the estimate obtained by Tucker and Thames (1983) from the F_e-plot of those data that is shown in Fig. 10a. This discrepancy seemed to us to be rather unseemly and so we compared the raw data (D, N) in Fig. 9 with the corresponding transformations (1/D, D/N) in Fig. 10a where D = D(0.50). We discovered that there is indeed an error in Fig. 10a at the point (1/D, D/N) for N=1. It can readily be seen that, for N=1, D(0.50) = 1.85 in Fig. 9 but 1.95 in Fig. 10a. The corrected transforms (1/D, D/N) of the data in Fig. 9 are shown in Fig. 10b. The curve of the regression model, $D^{-1} = \alpha_0 + \alpha_1 D/N$, together with the 0.95 CL are superimposed. The parameter estimates and 0.95 CL are $\bar{\alpha}_0$ = 0.075(0.029, 0.120) and $\hat{\alpha}_1$ = 0.240(0.196, 0.285) with $\overline{R}^2$ = 0.970. Examination of the regression diagnostics discloses that the observation at N=1 is both an influential point and an "outlier": Cook's D = 2.34, Studentized residual = 2.28.

The ratio $\bar{\alpha}_0/\bar{\alpha}_1$ = 0.310 is the corrected point estimate of α/β. Deletion of the observation at N=1 changes the ratio of the estimates to $\hat{\alpha}_0/\bar{\alpha}_1$ = 0.454, a change of nearly 50%. (The erroneous observation in Fig. 10a is shown as the filled symbol in Fig. 10b. If this observation replaces the correct one, the ratio of the estimates is $\hat{\alpha}_0/\hat{\alpha}_1$ = 0.423., as reported in Tucker and Thames.) The above sensitivity analysis illustrates the extreme fragility of the estimates of the parameters of regression models obtained from such small samples, i.e., in Fig. 10a the sample size is n=7 (i.e., 7 levels of N_i, T_i); a statistically adequate sample size is n = 20-40. However, it should also be remarked that the observation at N=1, which, as a high leverage point, dominates the estimates of α/β, also lies well beyond the limit of D/N = 10 Gy that was stipulated in Fowler 1984 to be the upper limit of the range of validity of the LQ model.

It should be clear that if one is going to use any point estimates of parameters (or functions thereof) in any consequential enterprise, then it is merely common prudence to insist upon only the most accurate estimate and then to further circumscribe it by appropriate diagnostics - and interval estimates as recommended by Park and Snee 1983. (See section 7.5 above.) Therefore, it should be noted that the re-analyses described in the present report and in Annexes II-IV included estimates of the bias, as well as the variance, of the estimates of $\underline{\beta}$ and of both linear and non-linear functions of $\underline{\beta}$, say $f(\underline{\beta})$ such as $\underline{x}^T\underline{\beta}$ and $\theta = \beta_1/\beta_2$, respectively.

As we remarked above, the only experiment described in Fig. 9 for which there are sufficient data to construct a dose-response curve is that for N = 2. This is shown in Fig. 11b, together with the 0.95 CL thereon. See below. The data at each of the other six levels of N are too weak to support the construction of the respective dose-response curves. However, all of these dose-response observations can be integrated by construction of a dose-response surface, $z = \beta_0$, $\beta_1 \log D + \beta_2 \log N + \beta_3 \log T$, on the full sample of n=19 observations shown in Fig. 9. Then dose-response curves, $z_i = \beta_0{}^* + \beta_1 \log D_i$, where $\beta_0{}^* = \beta_2 \log N_i + \beta_2 \log T_i$, together with 0.95 confidence limits, can be obtained for each of the seven levels of (N_i, T_i), $1 \leq i \leq 7$, as shown in Fig. 11b. It can be seen from Fig. 11b that the estimates of the slope and intercept of the dose-response curve 1 obtained from the dose-response surface are consistent with estimates obtained from the sample data at (N, T) = 2, 1). A moments' reflection will convince the reader that the family of curves obtained from the surface will also exhibit the decrease in slope with increase in N that is described in Fig. 9. (See also Fig. 6 of Tucker and Thames, 1983.) Figure 11b and the foregoing discussion provide an example of the "strengthening" of a set of weak experiments by a post-hoc salvage procedure in which each experiment "borrows strength" from all of the others - just as in a formal meta-analysis. (Figure 37 provides another example wherein the data of two studies with complementary weaknesses can be usefully pooled to strengthen the sample estimates of the parameters of a regression model: The gamma and neutron doses are highly correlated in both the

Hiroshima and Nagasaki LSS (T65D) data (BEIR III). However, the respective correlation structures are quite different. Hence, in the pooled data, the correlation between the neutron and gamma dose are much reduced. See section 1.1 above.)

It will be useful to illustrate the models of dose-response and isoeffect, and the duality relation between them for the case where the observed response has a Binomial distribution, B(π, n), where $0 \leq \pi \leq 1$ is the parameter, as in the case for the Tucker and Thames 1983 data of Fig. 9.

Important Topics:

Epistemology. Non-experiential data. Concordance of model and data. Consistency of a priori and sample information. Matrix calculus. Bias and variance of estimate. Model checking methods. Regression diagnostics. Residual. Hat matrix. Jackknife validation. RSS. PRESS. Covariance matrix. Row-deleted estimate. Multicollinearity. Ridge regression. Data augmentation. Matrix-weighted average. Mixed estimation. Delta method. Jackknife estimation. Mean shift outlier method. MELO estimates.

7.5.1 Binomial response. $0 < \pi < 1.0$. (Radiation Toxicity).

We will be interested in assessing and comparing two rival hypotheses, the LQ and Power-law, of the (causal) relation between the levels of a Binomial radiation response and the levels of the treatment variables dose, D, fractions N, and time T. The comparison will be based on the respective regression models of the two hypotheses constructed on common sets of data.

Figure 11a shows the non-linear transformation that maps the finite range of the observed response, the proportion π_i, $0 \leq \pi_i \leq 1$, at the i^th level of the treatment variables, into the infinite range that is required for the construction of a proper regression model:

$$\pi_i \longrightarrow z_i, \; -\infty < z < \infty, \; 1 \leq i \leq n.$$

Here z_i is the probit (or logit) transform of π_i:

$$z_i = \Phi^{-1}(\pi_i) \qquad \text{Probit}$$

$$z_i = \ln[\pi_i/(1-\pi_i] \qquad \text{Logit}$$

$\Phi(\;)$ is the standardized Normal distribution function, e.g., $\Phi(1.28) = 0.90$. The probit (or logit) transform is appropriate for the dose-response variable of Fig. 9 as shown in Fig. 11a.

7.5.1.1 Plausibility of rival hypotheses.

"More is different."

P. Anderson, 1972.

"... biological plausibility is the weakest kind of evidence for assessing cause-effect relationships. ... A lack of biological plausibility may indicate the limitations of our knowledge rather than the real lack of a causal association ..."

E. Neugebauer et al, 1987.

The two rival hypotheses, linear-quadratic and power-law, proceed from quite different philosophical bases. The LQ models are highly, one might even say excessively, reductionist. They are based upon the simple assumption that a specific radiation response of cultures of single cells in vitro - "reproductive death" - can be readily scaled to describe the radiation response of tissues in situ. For this reason the prior probability of the LQ model would seem to be low since, to borrow P. Anderson's rather cryptic phrase, "More is different": In biological systems, as in physical systems, the problem lies in the "... twin difficulties of scale and complexity." The radiation responses of large and complex aggregates of elementary entities are not likely, a priori, to be understood in terms of a simple extrapolation or scaling of a specific response of these entities. "Instead, at each level of complexity entirely new properties appear ..." And, again, as in physical systems, large and complex systems have many more degrees of freedom and consequently many more and different failure modes as well as couplings between these modes (Gilmore, 1992). Here, the whole is not only more than, but also very different from, the sum of its parts.

The model of the rival hypothesis is a homogeneous Power-law model of radiation response at the system level. It is a simple generalization of the phenomenological Weber-Fechner law of binary response - not a scaling of a putative mechanism. The LQ model is such a scaling - or an extrapolation. The distinction is non-trivial. Thus, in discrimination between these two rival hypotheses on a given set of observations of system response we feel that the prior odds ratio (Gilchrist, 1984; Leamer, 1978) must be in favor of the Power-law model.

But, of course, as the philosopher/statistician H. Kyburg (1970) has remarked, (echoing Osiander's preface to Copernicus' De Revolutionibis some four centuries earlier - see section 6 above.) "... it doesn't matter a damn bit where a hypothesis comes from. Any source, kooky or otherwise, is all right. Even statistical reduction may be such a source. Whatever the source, however, the hypothesis must fit the data." (emphasis added). For instance, what the philosopher/physicist Peirce (1901) has called the greatest piece of retroductive reasoning ever performed - Kepler's reductions of Tycho's observations to a set of confocal ellipses - had its "kookier" sources in Kepler's astrological and numerological commitments and practices by which the orbits of the six known planets were uniquely related to the five Platonic polyhedra. The actual choice of the Kepler hypothesis (ellipses) over the rival Ptolemy hypothesis (epicycles) seems to have been made on the criterion of parsimony (Ockham's Razor) since both models fit the data equally well - and the respective priors may be thought equally "kooky".

The modern Bayesian posterior odds ratio (See section 7.8) provides a formal method of systematically combining non-sample conceptual evidence, represented by a prior odds ratio, with empirical sample evidence represented by a likelihood ratio. And Ockham's Razor is still useful: "I say on the contrary, that the simplest law is chosen because it is the most likely to give correct predictions ..." (Jeffreys, 1961). It is also the case that the more parsimonious models will couple less of the random noise present in the sample observations into the estimated response. (Montgomery and Peck, 1982).

7.5.1.2 Models of Rival Hypotheses.

" The relationship of the overall treatment time and total dose to the outcome of a course of irradiation remains controversial more than 40 years after the first major attempt at a theoretical description."

M. Barton et al, 1992.

A) Dose-Response Models:

1) LQ + time. $z = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3 T$

2) Power-law. $z = \beta_0 + \beta_1 \log D + \beta_2 \log N + \beta_3 \log T$

z is the probit or logit transform of the proportion, π, of response.

B) Isoeffect Models:

1) LQ + time. $$D(\pi) = \frac{-\beta_1 + \sqrt{\beta_1^2 - 4(\beta_2/N)(\beta_3 T + \beta_0 - z)}}{2\beta_2/N}$$

2) Power-law. $\log D(\pi) = (z_\pi - \beta_0)/\beta_1 - (\beta_2/\beta_1)\log N - (\beta_3/\beta_1)\log T$

It should be noted that the dose-response models in A) are generalized linear models. That is, they may be written as $g(\mu_i) = \eta_i + \varepsilon_i$ where g(.) is the probit or logit transform of μ_i, the mean response, and ϵ_i is a random variable with mean zero and variance a function of μ_i. The linear predictor η_i is a linear function of the unknown parameter vector $\underline{\beta}$: $\eta_i = \underline{x}_i^T\underline{\beta}$ where $\underline{x}_i^T$ is the vector of predictor variables. On the other hand, the isoeffect model B1) is non-linear in the parameter vector whereas the model B2) is linear in the parameter vector $\underline{\alpha}$ where $\alpha_0 = (z_\pi - \beta_0)/\beta_1$, $\alpha_1 = \beta_2/\beta_1$ and $\alpha_2 = \beta_3/\beta_1$. Note that $\underline{\alpha}$ represents a re-parameterization of the original isoeffect equation. As Ratkowsky 1983, 1990 has observed, part of the non-linearity in a model (the so-called "parameter effects non-linearity") may "... often be reduced, sometimes drastically, by suitable re-parameterization." Moreover, "Note that transformations of parameters [in non-linear models] are very different from transformations of the responses [in generalized linear models]. Transformations of the response distort the response space and create a new expectation surface, thereby affecting the disturbances and the validity of the assumptions on the disturbances." D. Ratkowsky, 1990. In this context, "linear in the parameter vector $\underline{\alpha}$" means that the derivatives, $\partial \log D(\pi)/\partial\alpha_j$, $0 \leq j \leq 2$, are independent of α_j, whereas, "non-linear in the parameter vector $\underline{\beta}$" means that the derivatives, $\partial D(\pi)/\partial\beta_j$, $0 \leq j \leq 3$, are functions of β_j. (See also McCullagh and Nelder, 1989.) The distinction is important since for intrinsically non-linear models the maximum likelihood parameter estimates are biased, whereas for linear models these estimates are unbiased. And although the estimated response for a linearized model (a contingently non-linear model such as a power-law model, $y = ax^b$) may be a biased estimate of the response of the corresponding non-linear model, this bias is usually small - and bias-corrected estimates can be readily calculated (see Miller, 1984). If a comparison of rival hypotheses is to be made on the basis of the respective regression models, the two models must be commensurable. In the present comparison this means that we must compare the models of the isoeffect doses, $D(\pi) = D(0.50)$.

Figure 10c is a plot of a "deconstruction" of the data of Figs. 10a (and 10b) namely, the set of seven pairs of values of D(0.50) and fractions N obtained from Fig. 9. The superimposed curve is estimated from the non-linear isoeffect model of the LQ hypothesis constructed on these data. The dose-response model of the LQ hypothesis, $z = \beta_0 + \beta_1 D + \beta_2 D^2/N$, where z is the probit (or logit) transform has been reparameterized to give the cognate isoeffect model for $\pi = 0.50$, $z = 0$:

$$D(0.50) = \{\alpha_1 + [\alpha_1^2 - (4/N)\alpha_0]^{1/2}\}/(2/N)$$

where $\alpha_0 = (\beta_0/\beta_2)$ and $\alpha_1 = (\beta_1/\beta_2) = \alpha/\beta$. (N.B.: The correlation matrix of $\underline{\beta}$ disclosed that the model in β_0, β_1, and β_2 is over-parameterized (off-diagonal elements $\simeq$ 1.0.) This model of these data has $\bar{R}^2 = 0.989$ with parameter estimates (and standard error) as $\hat{\alpha}_0 = -3.866$ (0.483), $\hat{\alpha}_1 = 0.263$ (0.086). Here $\hat{\alpha}_1 = 0.253 = \alpha/\beta$ for the non-linear regression model, with 0.95 CL (0.032, 0.474) whereas the corrected F_e-plot of Fig. 10b gives the estimate $\alpha/\beta = 0.310$. It is the case, of course, that the Least Squares estimates of the parameters of non-linear regression equations are biased. However, it is also the case that the ratios of Least Squares estimates of regression parameters are biased estimates of the parameter ratio. This latter issue is discussed in sections 7.3 and 7.4. Moreover, "Inferential statements from non-linear regression lean heavily on normality and are accurate only for very large samples. In smaller samples, the accuracy of the large-sample results will vary greatly from problem to problem and can depend on the choice of parameterization. Standard errors produced using usual large-sample calculations that are given by most computer packages may be seriously in error in some problems and can either understate or overstate the precision of an estimate. As a first approximation, however, the standard errors can be used as they are used in linear regression. ... the ratio of an estimate to its standard error can provide a test statistic for the hypothesis that a parameter is equal to zero; approximate p-values are obtained from the standard normal distribution, not from a t distribution." S. Weisberg, 1985.

It will be noted that the shape of the distribution of the observations in the plot of Fig. 10c - concave-down - strongly suggests a Power-law isoeffect curve and therefore the logarithm of the data of Fig. 10c are re-plotted in Fig. 10d. It is immediately obvious that the data could be closely fit by the Power-law isoeffect model: $\log D(\pi) = \alpha_0 + \alpha_1 \log N$. This model of these data also has $\bar{R} = 0.989$ with parameter estimates and (standard errors) $\hat{\alpha}_0 = 0.288$ (0.015) and $\hat{\alpha}_1 = 0.413$ (0.018). This gives the model $D(0.50) = 1.925N^{0.413}$ krad. The shape of the distribution of the observations in Fig. 10c also suggests that a Power-law curve of the form $D(\pi) = \alpha_0 + \alpha_1\sqrt{N}$ might well fit these data and we found indeed that it did: $\bar{R}^2 = 0.990$. (It should be remarked that the non-linear isoeffect model of the LQ hypothesis also has, in fact, a fractional Power-law dependence on N.)

The excellent fits of the respective isoeffect models of the LQ and Power-law hypotheses might suggest that the cognate dose-response models are also valid for the van der Kogel data. However, as we will demonstrate below (in section 7.9.1) this is not the case for either model. Both dose-response models under-fit the dose-response data. Including a time factor log T improves the fit of both dose-response models. However, on the evidence of aggregate statistics of fit, the van der Kogel data still reject the thus augmented dose-response model of the LQ hypothesis. See Figs. 14-18 and the related discussions, but do not reject the cognate model of the Power-law hypothesis.

Although dose-response curves are drawn for each of the seven levels of N shown in Fig. 9 it should be recalled that only for N=2 are there sufficient data for statistically adequate (maximum likelihood) estimates of the form and parameters of a probit model to be constructed. The other curves must be (and in fact were - see Tucker and Thames, 1983) drawn by eye since the iterative maximum likelihood methods will not converge (at termination the curve describes a Heaviside, or step, function, i.e., the slope approaches infinity) if there is only one observation for which $0 < r_i/n_i < 1$ at a given level of N.

Figure 11b presents the dose-response curve (curve 0) and 0.95 confidence limits for the probit model constructed from the set of 4 observations at N = 2 in Fig. 9. This model is $z = \beta_0 + \beta_1 x_1$, where $z = \Phi^{-1}(\pi)$, $\Phi(.)$ is the Normal distribution function, i.e., $\Phi(1.28) = 0.90$, and $x_1 = \log D$. On the evidence of the Pearson χ^2 statistic on 2 df the model is consistent with those data. The evidence of Fig. 10d and Fig. 11b suggests that the causal model, $z = \beta_0 + \beta_1 x_1$, described by a dose-response curve, can be usefully generalized to the response surface model, $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where $x_2 = \log N$ and $x_3 = \log T$, of the van der Kogel data of Fig. 9. As remarked above, we have done so. The construction and testing of this Power-law model is described in Fig. 14 - Fig. 18 and section 7.9.1.

Figure 11b also includes two other estimates of the dose-response curve. These are curves 1 and 2. Curve 0 was estimated from the raw data, r_i/n_i at D_i, $1 \leq i \leq 4$, at N=2. Curve 1 is the trace of the dose-response surface for the Power-law hypothesis, $z = \beta_0 + \beta_1 \log D + \beta_2 \log N + \beta_3 \log T$, at N=2, T=1. Curve 2 is the dose-response curve constructed on the Berkson (1953) modification of the data at the two extreme responses, $r_i = 0$ and $r_i = n_i$ at N=2. Berkson (1953) replaces these with $(1/n_i)$ and $(1-1/2n_i)$, respectively. (An alternative modification gives the Bayesian estimates - obtained from Laplace's Law of Succession - of $1/(n_i+2)$ and $(n_i+1)/(n_i+2)$, respectively (Good, 1965).) For $n_i = 10$, the difference is not material: 0.05 vs 0.08 and 0.95 vs 0.92.

We have now examined several of the conceptual and statistical weaknesses of the F_e-plot. However, our examination has not been exhaustive and it will be useful to consider yet another one or two. Note that in Fig. 10a the estimate of α/E is obtained as the intercept of the straight line drawn on the scattergram. Such intercepts are invested, by all who construct such plots, with substantive meaning in the ratio α/β. However, it is a truism of regression analysis that the intercept of a straight line model, $y = a + bx$, can have substantive meaning only if both of two conditions are fulfilled (Younger, 1979):

1) It must be physically possible for the independent variable, x, to equal zero.

2) There must be observations in the region around x=0.

It can be seen at once that neither condition is fulfilled for the F_e-plot of Fig. 10a.

In the received practice of radiation biology, the computational - and conceptual - problems in the construction of a regression model presented by a non-Normal response, such as the Binomial, are solved by an "isoeffect transformation" such as shown in Fig. 10a instead of the statistically correct probit (or logit) transformation as described graphically in Fig. 11a. (See also Annex III, part 5 in which an "incidence transformation" was used for Poisson data.) Note that in such "isoeffect transformations" the number of predictor variables is always reduced by one, say

MONOTONIC TRANSFORMATION OF A PROBABILITY TO THE RANGE $(-\infty, \infty)$.

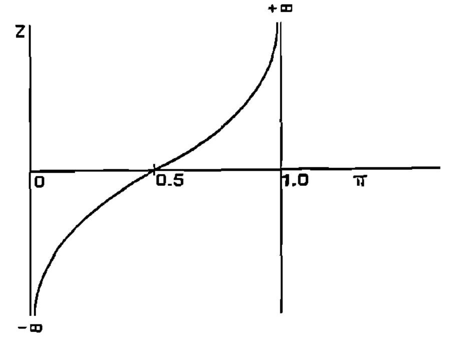


Fig. 11a. The plot describes the monotonic transformation that maps the interval (0,1) of the Binomial response variable π into the interval $(-\infty, \ \infty)$ of the probit transform, $z = \Phi^{-1}(\pi)$, where $\Phi(.)$ is the Normal distribution function, e.g., $z = 1.28$ for $\pi = 0.90$. In order to construct useful regression models of dose-response data the probit (or the cognate logit, or arc sine) transformation is the preferred transformation to apply to data in which the response has a Binomial distribution, rather than the isoeffect transformation, described in Fig. 10, since the probit transformation conserves the information on dose-response that is included in experiments such as described in Fig. 9 while the isoeffect transformation shown in Fig. 10 destroys most of such information. (Figure 9 includes information on the (conditional) response over the full range, $0 \le \pi \le 1$ while Fig. 10 includes information only on the (conditional) response at π · $= 0.50$.)

DOSE-RESPONSE CURVE AND 0.95 CL AT $(N_2, T_2) = (2,1)$.

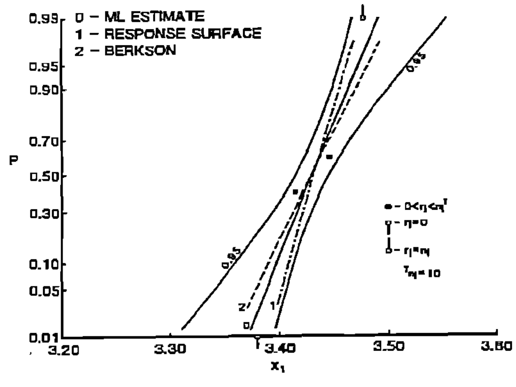


Fig. 11b. The curve labeled 0 is the trace of the dose-response curve for the probit model, $z_i = \eta_i = \underline{x}_i^T \underline{\beta} = \beta_0 + \beta_1 x_{1i}$, constructed from the data (N,T) = (2,1) in Fig. 9 where x_1 = logD. The 0.95 confidence limits are superimposed. Since the response has a Binomial distribution the appropriate generalized linear model is the probit. Note that it is only for (N,T) = (2,1) that there is enough information (4 levels of dose) to construct a dose-response model. None of the other six experiments shown in Fig. 9 include enough information to support a model.

The linear predictor, $\eta = \beta_0 + \beta_1 x_1$, of the dose-response curve can be readily generalized to $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ to give a dose-response surface where x_2 = logN, x_3 = logT. (See Fig. 18a.) Note that in this generalization, the slope β_1, of the dose-response curve is invariant, while the intercept, $\beta_0 \ = \beta_0 + \beta_2 x_2 + \beta_3 x_3$, is a function of (N,T). See Annex II, part 3. Note that invariance of a parameter, β_j, between sets of data obtained under different circumstances, is one criterion of a successful model: "What is to be the criterion for success in model-building? I shall adopt the proposition that we are successful to the extent that we can discern a pattern in the estimates of the parameters of the model, and that this implies the recognition of invariance. If we have several groups of data to which we fit straight lines and if it appears that a single slope parameter suffices for all, then we have established a degree of invariance over at least these sets of data." _ "The recognition of a subset of situations in which a degree of invariance occurs appears to be a crucial step in the development of a successful model" (Nelder, 1968). (See also Ehrenberg, 1975). Curve 1 is the trace of the response surface at (N,T) = (2,1).Curve 2 is the dose-response curve constructed on the data of Figure 9 at (N,T) = (2,1) modified according to Berkson's (1953) method for salvaging the information included in observations of extreme ($r_i = 0$ and $r_i = n_i$) binomial responses.

Note as well that the pooling of the data from the seven binomial dose-response experiments of Figure 9 each identified by a unique value of (N_i, T_i) is a successful instance of the salvage maneuver, data augmentation. (Another successful instance is described in Fig. 37 for observations in which the response has a Poisson distribution.)

Figure 11b should be compared with Fig. 12a which described the statistically adequate (generalized linear) model for data with a Poisson distribution of response.

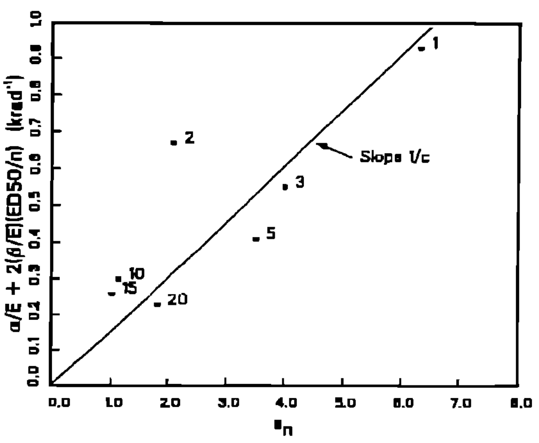


Fig. 11c. The figure is re-drawn from Fig. 8 of Tucker and Thames 1983. The number at each point is the fractionation, N. (See Fig. 10a.) It is obvious from the figure that the observations at N=2 and N=1 are, respectively, outlying and influential observations for this model. But, as we remarked above in Fig. 11b, the only one of the seven dose-response experiments described in Fig. 9 for which the experiment yielded sufficient data from which a dose-response curve could be constructed, by statistically adequate methods and criteria, was the experiment at N=2. And, although it is not acknowledged in Tucker and Thames (1983), it is equally obvious that the regression line has been forced to pass through the origin by imposing the constraint $(x^*, s_n) = (0,0)$ where $x^* = \alpha/E + 2(\beta/E)(ED50/n)$. But such a constraint is of dubious provenance: a) Since $\alpha/E \geq 0$ and $\beta/E \geq 0$, $x^* = 0$ requires that $ED50/n \leq 0$, which is impossible. b) $s_n = 0$ is also impossible since s_n is the slope of the dose-response at $p = 0.50$ for n fractions (see Fig. 9). But the slope of a dose-response curve is a maximum at $p = 0.50$ - as far from zero as it can get. It can be shown that the estimate of the slope, $1/c = 0.152$, reportedly obtained from this figure by Tucker and Thames (1983), is almost completely determined by the constraint at (0, 0) and the observation at N=1. But the observation at N=1 lies well beyond the range, 10 Gy, within which the LQ model is stipulated to be valid (Fowler, 1984, 1989), and it would seem that the constraint (0, 0) lies well beyond belief. Despite these obvious weaknesses in the methodology, Tucker and Thames 1983 subsequently deploy the estimate $1/c = 0.152$ with neither disclaimers nor qualification. The Fig. 11c provides a vivid illustration of a "strenuous and devoted attempt to force nature into the box [y = a + bx] provided by professional education" T. Kuhn, 1970.

from 3 to 2, these being the number of fractions, N, and the duration of time, T.

The isoeffect transformation suppresses information on a causal relationship, i.e., dose-response information, in order to extract a deterministic relationship, i.e., an isoeffect model. The "isoeffect transformation" was simply the first of a long line of systematic transmogrifications of dose-response data that have been deployed by the radiobiological community. This may represent an effort to achieve what theoretical physicist Mario Bunge (1979) has described as "... a single syntactic form that is invariant under a wide variety of semantic 'transformations'." Alternative, it may simply be what Thomas Kuhn has described as "... a strenuous and devoted attempt to force Nature into the box provided by professional education. The "form" (or "box") in the present case is the simple linear model, $y = a + bx$. In Tucker and Thames (1983) the F_e-plot, $D^{-1} = a + bD/N$, is only one example. Figure 11c describes a still more "strenuous and devoted attempt." Figure 11c is re-drawn from Fig. 8 of Tucker and Thames (1983). The number at each point is the fractionation, N. (See Fig. 10a.) It is obvious from the figure that the observations at $N = 2$ and $N = 1$ are, respectively, outlying and influential observations. And, although it is not acknowledged in Tucker and Thames (1983), it is equally obvious that the regression line has been forced to pass through the origin by imposing the constraint $(x^*, s_n) = (0, 0)$ where $x^* = \alpha/E + 2(\beta/E)(ED50/n)$. But such a constraint is surely of dubious provenance: a) Since $\alpha/E \geq 0$ and $\beta/E \geq 0$, $x \equiv 0$ requires that $ED50/n \leq 0$, which seems unlikely. b) $s_n \equiv 0$ is also unlikely since s_n is the slope of the dose-response at $p = 0.50$ for n fractions (see Fig. 9). But the slope of a dose-response curve is a maximum at $p = 0.50$ - as far from zero as it can get. Despite these obvious weaknesses in the methodology, Tucker and Thames (1983) subsequently deploy the estimate $1/c = 0.152$ without any disclaimers or qualifications.

More recent examples of forcing Nature into the "box" $y = a + bx$ can be found in the papers of Travis and Tucker, 1987, Withers et al, 1988, and Fowler, 1991. In Travis and Tucker, the form appears as $D(\alpha/\beta + D/N) = a + bT$. In Withers et al it is $TCD50 = a + bT$. In Fowler 1991 it is $D[1 + (D/N)/(\alpha/\beta)] = a + bT$. In all of the foregoing versions of the basic form $y = a + bx$, the estimates of α/β are a "given", i.e., they are a priori, or non-sample, estimates. (It is unfortunately the case that such estimates are invariably "given" without any concomitant estimates of either uncertainty or irrelevance.)

In Tucker and Thames 1983, Travis and Tucker 1987, and Fowler 1991, the estimates of a and b of the basic linear "box" are obtained by regression methods. However, in none of the three studies is there a proper regression model. a) In Tucker and Thames the dependent variable, D(0.50), appears on both sides of the equation. In Travis and Tucker and in Fowler, an ad hoc, a priori, weighted linear combination of two dependent variables ($D(\pi)$ and N) is regressed on a single predictor variable, T. But in the proper version of the cognate (multiple) regression model, a dependent variable, $D(\pi)$, is regressed on a linear combination of two predictor variables, N and T, in which the weights are determined from the sample. In Withers et al, the TCD50 is an ad hoc combination of $D(\pi)$ and N. However, the validity of the relation between the TCD50 and T that is presumed to subsist in the data of their Fig. 1 (see our Fig. 47) and the sample estimate of the slope is assessed and obtained by the transcendently ad hoc procedure of "eye-balling it." Again, such a transformation also collapses a causal model of dose-response into a deterministic model of isoeffect.

Since N and T are (usually) highly correlated, in both experimental and non-experimental studies, e.g., $N = (5/7)T$ in most clinical data, the resulting isoeffect equation is effectively of the (now familiar) simple linear form $y = a + bx$, with $y = f\{D(\pi)\}$ and either $x = g(T)$ or $h(N)$ where $f\{.\}$, $g(.)$, and $h(.)$ are arbitrary functions, e.g., for the old "cube root" law, $f\{D(\pi)\} = \log D(\pi$ and $g(T) = \log T$. Indeed, such was the basic form from which the NSD was derived.

If N and T are highly correlated, one of them may contribute little additional information on the observed conditional dose-response. As a consequence, it is usually found that say, $\bar{\beta}_3/\sqrt{Var(\hat{\beta}_3)} < 1$, and often $\bar{\beta}_3 < 0$ as well, in dose-response models of clinical data. Of course, the high correlation of N and T in non-experimental, clinical, data is required by the ethical - and legal - constraints to adhere to the "standard of practice in the community": treatments are given MTWThF. However, the presence of similar levels of correlation in N and T in designed animal experiments is a consequence of flawed design which seems to be a result of one of the ontological weaknesses discussed above - the failure to correctly take into account the multivariate nature of the deterministic part of the response.

7.5.2 Poisson response. $0 < m < \infty$. (Radiation Lethality).

"To specify a Poisson regression model, it is assumed that the dependent variable follows a Poisson distribution and that a rate function, $\lambda(X, \beta)$, that describes the relationship between disease rates, the predictor variables (X), and the unknown vector of parameters (β) is given."

E. Frome and H. Checkoway, 1985

"The slope of each plot is the quadratic inactivation constant (β) and the intercept with the zero dose axis is the linear inactivation constant(α)."

T. Alper, 1980

Let us consider next the appropriate generalized linear model for a response with a Poisson distribution, $P(\lambda)$, with parameter $\lambda = mc$, where c is a number of observation units, e.g., 10^5 person-years, 10^3 cells, etc.,and m is a rate function at dose D. For the LQ hypothesis we have the dose-response model,

$$m = \exp(\beta_0 + \beta_1 D + \beta_2 D^2) \qquad (1)$$

The rival model is the so-called target theory model. There are several versions of this model. In this report we have found it sufficient to consider only the so-called single-hit, multi-target model,

$$m = \beta_0\{1 - [\exp(\beta_1 D)]^{\beta}2\}$$

We defer a direct comparison of these two models on a common data set to sections 7.9 and 7.10. However, we now examine the LQ model - and received practices - more closely. The "survival transformation" for the LQ model is,

$$me^{-\beta}0 = \exp(\beta_1 D + \beta_2 D^2) = S$$

where $e^{\beta}0 = m_1$ is the response at zero dose, $D_1 = 0$ and $0 \leq S \leq 1$. A subsequent log transformation gives the familiar "cell-survival equation,"

$$\log S = \beta_1 D + \beta_2 D^2 \qquad (2)$$

There are two alternative transformations that further reduce the equation to the simple straight line form, $y = a + bx$, for which estimates of the slope and intercept can be obtained either by graphical procedures or by ordinary least squares methods. Which of these transformations is used depends upon the number of fractions, N, into which the dose is divided:

1) $N \geq 1$. F_e-plot (Douglas and Fowler, 1976)

This requires two additional sets of assumptions. The first set of assumptions is

a) That the survival following N equal fractions of size d can be represented as, $SN = \Pi S_1 = \exp(N(\beta_1 d + \beta_2 d^2)) = \exp(\beta_1 D + \beta_2 D^2/N) = S$. This assumption implies that, "... the same level of killing results from each successive [equal] dose ...", that, "... proliferation of survivors is negligible" and that the cells are homogeneous in their radiation response. Taking the logarithms, we have

$$\log S = \beta_1 D + \beta_2 D^2/N \qquad \text{dose-response model} \qquad (3)$$

However, this last assumption requires that the biological effect of a given increment of dose is the same at every level of dose but, as remarked in Annex II, part 3, this is inconsistent with a biological law of wide generality which requires that the response of a system exposed to repeated injury change both quantitatively and qualitatively. For this reason the response of a biological system to a given increment of dose at N=1 may be quite different from the response to that dose at N > 1. Thus, on this criterion, the data described in Fig. 9 are heterogeneous. (See also Figs. 2a, 10, and 22.) Note that in the probit model of the Power-law hypothesis of a quantal response, the response is a function of the logarithm of the dose (Weber-Fechner law) whereas in the multifraction LQ model of a quantal response (3) the logarithm of the response is a function of the dose. Note also that in the probit model with linear predictor, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where $x_1 = \log D$, $x_2 = \log N$, and $x_3 = \log T$, the covariate $x_2 = \log N$ is a confounder whereas, in the LQ predictor, with $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N$, the covariate N is an effect modifier - D(D/N) is an interaction term. That is, the covariate enters in different roles for the two rival models of the Binomial response.

Returning now to the Poisson response of equation (3) and dividing equation (3) by log S and by D we have the equation of the F_e-plot of Fig. 10:

$$D^{-1} = (\beta_1/\log S) + (\beta_2/\log S)D/N \qquad \text{isoeffect model} \qquad (4)$$

The second set of assumptions required in the F_e-plot is

b) That a unique level of S(urvival) of some usually unidentified (and nearly always unidentifiable) cell population corresponds, somehow, to an observed level, π, $0 \leq \pi \leq 1$, of a binary tissue or organ response which is a proxy for it; it is assumed that, "isoeffect means isosurvival": Note that the "survival transformation", $m_i \longrightarrow S_i$, $1 \leq i \leq n$, maps a variable with a (semi-) infinite range $(0, \infty)$ into a variable with a finite range: $0 \leq S_i \leq 1.0$. The assumption that "isoeffect means isosurvival" then conflates the Poisson and Binomial distributions. However, this conflation is only an artifact, not a natural law. The assumption has not been demonstrated to be valid in any of the cases in which it is made. It is, perhaps, a plausible conjecture but it does not appear to be a demonstrable one. Thus, in the multifraction LQ model the dose $D = D(S)$ is now an isoeffect dose, $D = D(\pi)$. The usual level of tissue response that is chosen is $\pi = 0.50$. Note that, again, the isoeffect equation has the simple form $y = a + bx$ where $y = D^{-1}(\pi)$ and $x = D(\pi)/N$. See Fig. 2a.

2) N=1. Cell Inactivation Plot (Chapman, 1978).

Dividing equation (2) by D we have the equation of the Chapman plot of Fig. 2b:

$$(\log S)D^{-1} = \beta_1 + \beta_2 D \qquad (5)$$

In this case it is the dose-response equation that takes the form $y = a + bx$ where $y = D^{-1}\log S$ and $x = D$. As described by Alper (1980) and Fowler (1984) $\hat{\beta}_1$ is an estimate of α and $\hat{\beta}_2$ is an estimate of β. Note that in the case of the Chapman reciprocal dose plot of Fig. 2b the levels of survival, S, are known whereas, this is not the case for the F_e-plot of Fig. 10a. However, neither the Chapman plot, nor the F_e-plot describe proper equations since in both equations the dose, D, appears on both sides of the equality sign.

It is of interest to compare the estimates of α, β, and α/β obtained from the Chapman plot of Fig. 2b with the cognate estimates from the appropriate generalized linear model (Poisson) of the same data. For the Chapman plot we have $\bar{\alpha} = -0.304$, $\beta= -0.098$, $\hat{\alpha}/\hat{\beta} = 3.112$. For the Poisson model we have $\hat{\alpha} = -0.436$, $\hat{\beta} = 0.074$, and $\bar{\alpha}/\hat{\beta} = 5.892$. (See Annex I, part 5 and Annex IV, part 6.) The respective differences are obviously non-trivial.

There are several comments to be made. First, the devolution of the Poisson log-linear dose-response model of equation (1) into the cell-survival model of equation (2) can be shown to be equivalent to imposing an arbitrary a priori constraint on the model parameter vector: $\beta_0 = \log m_1$. That is, the dose-response curve is constrained to pass through the level of response, m_1, that is observed at $D = D_1 = 0$. Compare Figs. 12a and 12b. But then this maneuver effectively assigns, a priori, an infinite weight to this observation: $w_1 \longrightarrow \infty$ ($10^6 \leq w_1 \leq 10^{10}$, say). Or, equivalently, in terms of the covariance matrix, ψ, of the prior constraint, $\beta_0 = \log m_1$, we have $10^{-6} \leq \psi \leq 10^{-10}$. That is, the ordinary least squares estimates of the parameters $\beta_1(=\alpha)$ and $\beta_2(=\beta)$ obtained from the survival model are equivalent to constrained least squares estimates obtained from the dose-response model in equation B(1). But there would seem to be no empirical warrant for the assumption that $w_1 \longrightarrow \infty$ (or that $\psi \longrightarrow 0$). It can be shown that, as would be expected intuitively, the proportion of the sample information on dose-response that is represented by the (constrained) survival model is only $\theta_s = 0.67$. Obviously, imposition of the survival constraint degrades the "fit" of the model to the sample data: the residual sum of squares is inflated, ($RSS^{**} = RSS + (\hat{\underline{\beta}} - \hat{\underline{\beta}}^{**})^T X^T X(\hat{\underline{\beta}} - \hat{\underline{\beta}}^{**})$, in terms of a Normal theory model). In the case of the Chapman model of the LQ hypothesis in equation (5) it can be further shown that the ordinary least squares estimates of β_1 and β_2 correspond to weighted least squares estimates $\hat{w}_1 = 10^6$, $\bar{w}_i = D_i^{-2}$, $2 \leq i \leq n$. See Fig. 2b and Annex IV, part 6.

The further devolution of the dose-response (cell-survival) model of equation (1) into the multifraction, $N \geq 1$, isoeffect model of the F_e-plot of equation (4) and (5) results, as we remarked

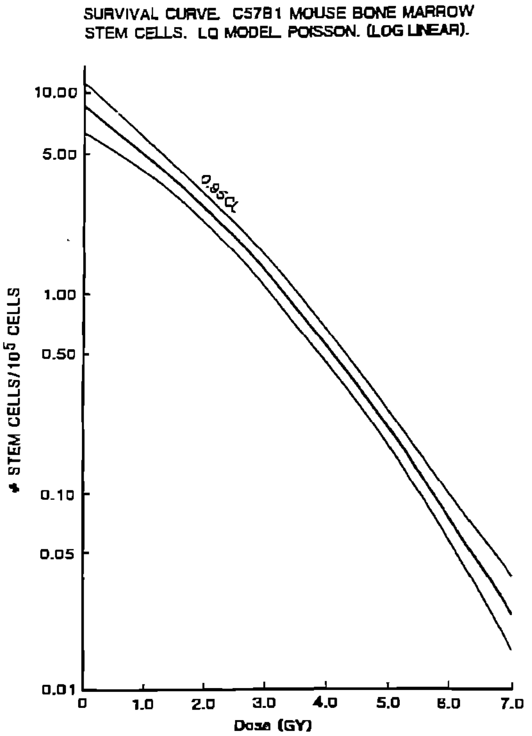


Fig. 12a. Plot of the dose-response curve and 0.95 confidence limits for the LQ model of mouse bone-marrow stem-cell survival with the parameter vector estimated by IRLS methods under the assumption that the radiation response has a conditional Poisson distribution.

Under this assumption the appropriate generalized linear model of the LQ hypothesis for cell survival data is the Poisson log-linear, $m_i = \exp(x_i^{2}\beta)$, where m_i is the Poisson rate parameter $\exp(x_i^{T}\beta)$ and $x_i^{T}\beta = \beta_0 + \beta_1 D + \beta_2 D^2$. Compare with Fig. 11b which provides the cognate plot for the generalized linear model of a radiation response which has a conditional Binomial distribution.

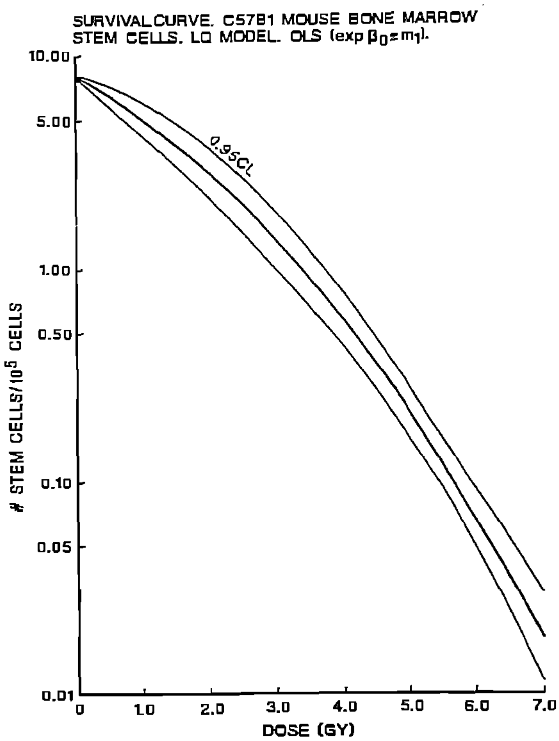


Fig. 12b. Plot of the dose-response curve for the LQ model of mouse bone-marrow stem cell survival, $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$, with the parameter vector estimated by ordinary least squares methods (OLS) under the constraint $\exp(\beta_0) = m_1$, where m_1 is the observed response at $D_1 = 0$, and the assumption that the response has a conditional Normal distribution. The estimate of β is the Mixed, or posterior, estimate obtained by including the constraint as non-stochastic prior information on β:

$$r = R\beta + v;\ E(v) = 0,\ \mathrm{Var}(v) = \psi$$

where $r = \log m_1$, $R = (1,0,0)$, $\psi = [0]$, the null matrix. But such a priori information would appear to be clearly fictive. It is the case, of course, that this estimate of β is equivalent to that obtained with the survival transformation applied to the data: $m_i \longrightarrow S_i = m_i/m_1$, $1 \leq i \leq n$ (a "relative risk" model of mortality). This is analogous to the incidence transformation $m_i \longrightarrow I_i = m_i - m_1$ (an "absolute risk" model of incidence) for the LQ model of mutagenesis $m_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$, described in Figs. 7 and 8.

Note that the effect of imposing the constraint, $\exp(\beta_0) = m_1$, constricts the confidence limits to zero at $D_1 = 0$. It also degrades the fit of the model to the data, of course, since it can be shown that the sum of squared residuals under the constraint, RSS^{**}, exceeds the unconstrained sum RSS:

$$RSS^{**} = RSS + (\hat{\beta} - \bar{\beta}^{**})X^{T}X(\bar{\beta} - \bar{\beta}^{**}).$$

It can be shown that, as might be expected intuitively, the proportion of information in the constrained estimates of β_1 and β_2 that is contributed by the constraint is $\theta_p = 0.333$.

Since the range of the data includes observations at D = 0, then β_0 can be interpreted as the mean of the log response at D = 0. The survival transformation, $S_i \longrightarrow m_i/m_1$, which is equivalent to the constraint $\ln m_1 = \beta_0$, implies that the single observed response, $\ln m_1$, is used as the estimate of this mean response. Such a practice is obviously very difficult to defend.

This survival transformation is also equivalent to a weighted least squares procedure in which the observation (m_1, D_1) is assigned, arbitrarily, an infinite weight, $w_1 \longrightarrow \infty$ (say $w_1 = 10^6$), and the observations (m_i, D_i) are assigned $w_i = 1$, $2 \leq i \leq n$.

The differences between the respective estimates of β_1, β_2, and β_1/β_2 ($=\alpha/\beta$) for the LQ model that are obtained from the two assumptions, Poisson (IRLS) (Fig. 12a) and Normal (CLS) (Fig. 12b) are substantial. For example

	IRLS	CLS	Ratio
β_1	-0.469	-0.362	1.296
β_2	-0.053	-0.072	0.736
β_1/β_2	8.902	5.020	1.773

above, in only a kind of pseudo-model for, "Usually a model is designed to explain relationships that exist among quantities which can be measured independently in an experiment; these are the variables of the model" (Bard, 1974). It is obvious that the left-hand side of equation (4) (or equation (5)) cannot be "measured independently" of the right-hand side of the equation.

Moreover, for the F_e-plot the levels of survival S that correspond to the levels of observed binary response in a biological system are often unknown (Tucker and Thames, 1983). Indeed the cell population whose survival is at risk often cannot be identified - or even defined. Therefore, the usual multi-fraction LQ model of functional defects in a biological system is not only value-free, but also content-free. Since the response is specified only as $E = -\ln S$ the LQ model has no empirical content. (But, "... the Indistinctness of Ideas, ... was long one main impediment to the progress of science in the middle ages ..." W. Whewell, 1856). Because log S is unknown only the transforms $\beta_1/\log S = \beta_1^*$ and $\beta_2/\log S = \beta_2^*$ can be estimated from the data. Although sample estimates of the parameters β_1 and β_2 cannot be obtained, the ratio $\alpha/\beta = \beta_1/\beta_2$ is regularly reported in the literature. Note, also, that the isoeffect model that is cognate to the dose-response model, $\log S = \beta_1 D + \beta_2 D^2/N$, is not $D^{-1} = (\beta_1/\log S) + (\beta_1/\log S)(D/N)$ of the Fe-plot but is rather the positive root of the quadratic dose-response equation:

$$D = \{-\beta_1 + [\beta_1^2 - 4(\beta_2/N)\log S]^{1/2}\}/2(\beta_2/N)$$

where $D = D(S)$. That is, it is a non-linear regression model not the linear regression model of the F_e-plot. Since S is usually unknown as well, as remarked above, the non-linear equation presents some rather refractory problems in estimation as well as the difficulties of interpretation posed by that lack of empirical content of the equation that was remarked above. But note that if the LQ isoeffect model is derived from the appropriate generalized linear model of the LQ hypothesis as describes in B above, namely, $D(0.50) = \{-\beta_1 + [\beta_1^2 - 4\beta_2\beta_0/N]^{1/2}\}/(2\beta_2/N)$, the difficulty created by the unknown value of the level of survival, S, disappears. (In point of fact, as we have remarked, the above equation for $D(\pi)$ is over-parameterized, and the re-parameterized version, $D(\pi) = \{-\alpha_1 + [\alpha_1^2 - 4\alpha_0/N]^{1/2}\}/(2/N)$, is required in order to achieve stable parameter estimates. Note that $\alpha_1 = \alpha/\beta$.)

It is of no little interest to recall that for its first ten years of its public life the received multifraction $N \geq 1$ LQ model has omitted a variable, T, the duration of time in which the N fractions are delivered, and which has been shown in many studies to have an important effect on the observed level of response (or else is a good proxy for another variable that does have an important effect), and which, also can, of course, be readily measured, but has included an occult variable, S, which cannot be measured, since it refers to an unobserved (and unobservable) level of response of an unknown (and unknowable - or usually even undefinable) target cell population. The fact that the variable T, or any function thereof, is omitted from the LQ model suggests that it may underfit some data leading to biased (aliased) estimates of the parameters. We will briefly examine some of the more recent developments in the evolution of the LQ model later in this report.

7.5.3 Models of low dose and high dose radiation effects

Radiation effects in organisms can be usefully dichotomized as stochastic and non-stochastic effects. The present report has been concerned only with published studies of former class in which the frequency of occurrence, but not the severity, of the effect is a function of dose. There are two single-parameter distributions of frequency of occurrence which are relevant: the Binomial, parameter π, and the Poisson, parameter λ. For small values of π, say $\pi < 0.1$, the Poisson may provide a satisfactory approximation to the Binomial.

For our immediate purposes it is convenient to further dichotomize stochastic radiation effects into low-dose effects and high-dose effects. The distinction between the two classes of dose levels is not sharp; for present purposes we take $0 \leq D \leq 10$ Gy as defining low-dose effects and $10 \leq D \leq 60$ Gy as high-dose effects. We shall find both similarities and differences in the several weaknesses of the published studies of the respective groups.

7.5.3.1 Models of biological effects of ionizing radiation at low doses

It appears from the published reports that low-dose effects are most often represented by dose-response models. There does not appear to be any scientific interest in isoeffect models for the low dose effects that we consider here. In the present review two of the classes of low-dose effects of interest are a) mutagenesis; b) cell-killing. In both, the response can be adequately modelled by (conditional) Poisson distributions. The respective Poisson models of the received LQ hypothesis for a sample of n observations of each response are the Poisson linear and Poisson log linear models:

a) $m_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$. $1 \leq i \leq n$. $0 \leq m_i \leq \infty$.

b) $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$. $1 \leq i \leq n$. $0 \leq m_i \leq \infty$.

where m_i is the (conditional) Poisson rate parameter at dose D_i. It is a common maneuver in received practice in modelling both responses to fix the value of β_0 at the level of response, m_1, at $D_1 = 0$. For the models of the effects a) and b), respectively, this is achieved by constructing models of the net incidence (excess risk) $I_i = (m_i - m_1)$, for which $\beta_0 \equiv m_1$, and survival (relative risk) $S_i = (m_i/m_1)$, for which $\beta_0 \equiv \log m_1$, respectively. There are several implications of this practice. We consider only two and base our expositions on the cell-survival model; the generalization to the mutagenesis (and, of course, carcinogenesis) models is obvious. 1) The transformation of the n observations, (m_i, D_i) to (S_i, D_i), where $S_i = m_i/m_1$, $1 \leq i \leq n$, can be shown (see Annex IV, part 6) to be equivalent either to imposing (arbitrarily) a specific version of the general linear constraint, $\underline{r} = R\underline{\beta} + \underline{v}$, where $E(\underline{v}) = 0$, and $Var(\underline{v}) = \psi$: $\underline{r} = \log m_1$, $R = (1, 0, 0)$, $\psi = [0]$, the null matrix, or, to arbitrarily assigning a set of weights, $w_1 = 10^6$, $w_i = 1$, $2 \leq i \leq n$, to the n observations. (See Annex II, part 5.) The constrained estimates of the parameters β_1 and β_2 of the model thus obtained are biased with respect to the unconstrained estimates. For the LQ model of the mouse bone marrow stem cell survival data, the constrained estimates of α/β are less than the unconstrained estimates by a factor of 0.56. Moreover, the imposition of the constraint inflates the residual sum of squares, RSS; i.e., degrades the "fit" of the model to the data. However, the constrained estimates are of speciously high precision. (See the discussion of Mixed estimation in section 7.2.4.) This specious inflation of the precision of the estimates can be seen most vividly in the peculiar shape of the 0.95 confidence limits on the estimates response presented in Fig. 12b. Following the "survival transformation" only 2 parameters, β_1 and β_2, remain to be estimated from the data by formal statistical methods. 2) Since the range of the data includes observations at D=0 then β_0 can be interpreted as the mean of the (log) response at D=0. The survival transformation, m_i/m_1 S_1, which is equivalent to the constraint $\ln(m_1) = \beta_0$, implies that the single observed response, $\ln(m_1)$, is used as the estimate of the mean response at zero dose. However, it is obvious that a better (smaller bias, larger precision) estimate of the response at $D = 0$, namely $\exp(\hat{\beta}_0)$, can be obtained by using the information in all sample observations $(y_i, \underline{x}_i^T)$, $1 \leq i \leq n$, rather than in only the single observation $(y_1, \underline{x}_1^T)$. Both implications serve to demonstrate that the current practices of transforming the set of observed responses to differences, $I_i = m_i - m_1$, and ratios, $S_i = m_i/m_1$, prior to fitting a model, cannot be defended. (N.B.: It is of interest to remark that the "survival transformation" and the "isoeffect transformation" reduce the number of parameters to be estimated from the data by quite different procedures. In the former it is achieved by constraint, e.g., $\beta_0 \equiv \log m_1$, and in the latter by re-parameterization, $\alpha_j = \beta_j/\beta_1$, $j = 0,2,3$. Reducing the number of parameters, of course, will always increase both the precision and bias of the parameter estimates.)

The parameter estimates for the models of the low dose effects which are published in the literature are obtained by non-weighted non-linear least squares in the case of the mutagenesis data, that is, the power-Normal models described in Figs. 5 and 6, and non-weighted linear least squares in the case of the cell-survival data. (See Annex III, part 5, and Annex IV, part 6, respectively.) Thus, in the received practice, the information on the nature of the distribution - Poisson - of the random part of the response, and hence of the intrinsic weight, w_i, $1 \leq i \leq n$, of each observation, that is provided by the sample observations is ignored in obtaining the parameter estimates of each model that are published in the peer-reviewed literature.

For the purposes of the present report there is also a third class of low-dose effect that is of interest - carcinogenesis (and leukemogenesis) in humans and mammary neoplasia in rats. Both responses are typically described in the literature by dose-response - not isoeffect - models. In the former response the incidence, or mortality, rate m_i can be satisfactorily represented by Poisson linear models since $0 \leq m_i \leq \infty$ and the incidence (and mortality) rates are quite small. However, in the latter case, the incidence rates are quite large - approaching 100%. The response will have a Binomial distribution and therefore a probit (or logit) model of dose-response is required: $z = \beta_0 + \beta_1 x_1$, where $z = \Phi^{-1}(\pi)$, $0 \leq \pi \leq 1$, $x_1 = \log D$ and $\Phi(\ )$ is the Normal distribution function. For the published studies of these models, which we have reviewed, statistical weaknesses of a different sort are present.

In the BEIR III (NAS/NRC, 1980) study on radiocarcinogenesis and leukemogenesis in the LSS sample that we re-analyzed (see also Herbert 1983, 1986c, 1989c), the correct generalized linear models in which link functions, g(.), that were consistent with the distribution - Poisson - of the stochastic part of the response were deployed. However, in the published study the deterministic part of the received Poisson model, LQ-L, of the LSS leukemia incidence data is misspecified owing to an over-reading of the Pearson chi-squared goodness of fit measure: It can be seen in Table V-8 of the BEIR III (1980) report that the difference in the respective values of the Pearson chi-squared statistics, RSS = $\Sigma\chi_i^2$, (where χ_i, $1 \leq i \leq 16$ is the chi-squared residual) for the LQ-L model vs the L-L model (and the LQ-L model vs the Q-L) are less than 3.84, the value of the 0.95 quantile of the Pearson chi-squared distribution for one degree of freedom; hence the LQ-L model overfits the LSS data. This overfitting must, of course, as discussed above in sections 7.3 and 7.4, lead to estimates of the so-called "cross-over dose" in which the bias is very large indeed. See also Annex III, part 4.

We also note two of the more curious general features of the group of models proposed in the BEIR III (NAS/NRC, 1980) report:

1) The respective estimates of the covariance matrices Var($\hat{\beta}$) of the parameter estimates, $\hat{\beta}$, for models of the rates of leukemia incidence (Table V-9) non-leukemia mortality (Table V-9) and breast cancer incidence are each incorrectly multiplied by the respective heterogeneity factors, $h = \chi_c^2/(n-k)$, where χ_c^2 is the chi-squared goodness-of-fit statistic and (n-k) is the number of degrees of freedom for the model. For each of the models of the respective responses, save for the LQ-L and L-L models of breast cancer incidence, $0.97 \leq h \leq 1.26$. However, for the LQ-L and L-L models of breast cancer incidence $h \sim 0.52$. Thus, in the case of the L-L model, the model of choice of the BEIR III report, vide infra, the variance of each of the parameter estimates is spuriously deflated by a factor of – 0.70. (See also Herbert, 1986c.)

2) The second curious feature of the group of BEIR III models is that each includes a (0,1) index variable to identify the site of the observations, Hiroshima or Nagasaki. However, for the models of leukemia incidence and non-leukemia cancer mortality the estimate of the coefficient of the index variable is less than its standard error. But it is good practice, on a minimum mean-squared error criterion, to omit such variables from the model since their presence inflates the variance of the estimates of the remaining parameters - and functions thereof - more than their absence inflates the bias of these estimates. (See also Herbert, 1986c.)

In the case of the rat mammary neoplasia studies that we re-analyzed, the information on the nature of the distribution, Binomial, of the random part of the response that was provided by the sample observations (r_i responders in n_i at risk at dose D_i, $1 \leq i \leq n$) was ignored by the original investigators (Bond et al, 1960; Shellabarger et al, 1969). As a consequence, in these two published studies, the (received) dose-response model that is deployed is the linear model, $\pi = \alpha_0 + \alpha_1 D$. But for this model the distribution of tolerance dose is rectangular, which is wholly inappropriate. see Dobson (1983). For a Binomial response, the correct model is the probit, $z = \beta_0 + \beta_1 x_1$, where $z = \Phi^{-1}(\ )$ and $x_1 = \log D$. For the probit model the distribution of tolerance dose is Normal. See Dobson (1983).

In all of the published studies on mammary neoplasia that we reviewed, the form of the distribution of the random part of the response and the form of the deterministic part of the

The 0.95 confidence limits on α_0 for the model of the reduced (n=5) data include 0 which implies the model, $x = \alpha_1 D$. But, as in the case of the mutagenesis data, this is inconsistent with prior information on the level of spontaneous incidence. It appears the received Normal theory model, $x = \alpha_0 + \alpha_1 D$, deployed in Shellabarger et al (1969), achieved the misspecification of both the deterministic and random parts of the response. See Figs. 30 and 33.

Note that in these data on low-dose radiation effects in which the response has a Binomial distribution, and thus requires a probit model, the numbers at risk at each treatment level are sufficiently large, $n_i > 25$, so that the sampling distribution of $\chi^2 = \Sigma \chi_i^2$, where χ_i is the ith chi-residual, is well-approximated by the Pearson chi-squared distribution.

On the other hand, the data of Fig. 9 on high-dose radiation effects also have a Binomial distribution but the numbers at risk are so small, $n_i < 10$, that the sampling distribution of χ^2 is not well-approximated by the Pearson chi-squared distribution. The respective numbers at risk appear to be typical for the two areas of investigation: low dose effects and high dose effects.

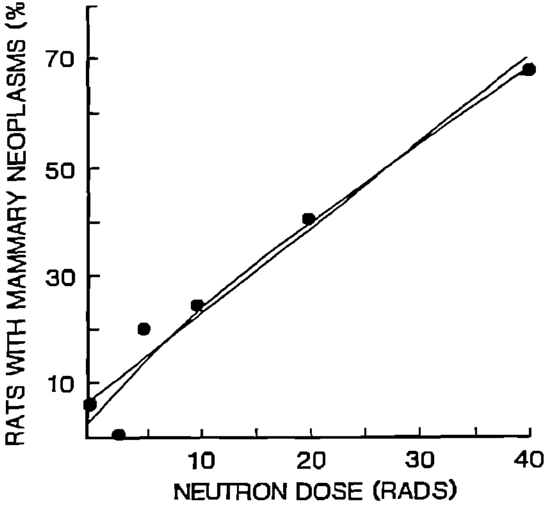


Fig. 13b. Plot of the Montour et al (1977) data on incidence of mammary neoplasia vs neutron dose for female Sprague-Dawley rats. Note that many of the weaknesses in the design of the Shellabarger et al (1969) and Bond et al 1960 experiments described in Fig. 13a are repeated in this design. This invariance in the design represents a "propagation of error" across a time-span of nearly a generation. Note that for a log dose metameter the observations are more uniformly distributed and the estimates no longer dominated by the observation at the highest dose. (See Figs. 30 and 32.)

by Fieller's theorem are adequate: $\hat{\theta}$ = 8.64; 0.96 CL = (5.30, 13.84).

Montour et al 1977 reported their analysis of the same two data sets using models of dose-response in which both the functional form of the deterministic part of the response and the distributional form of the random part were mis-specified: $\pi = \alpha_0 + \alpha_1 D$. Based on the ratio of the slopes of the respective linear models of the γ-ray and neutron data they reported (only) a point estimate of $\hat{\theta} = 4.3$. On the basis of a curvilinear model of the θ-ray response they also reported that the RBE increased with decreasing dose. However, our secondary analyses of the Shellabarger et al, 1969 and Montour et al, 1977 data sets that are reported in Annex III, part 6, disclose that, on the basis of the probit models, the neutron RBE is independent of dose, since the respective probit dose-response curves are parallel. This is contrary to the conclusions of the published studies - and to received opinion on the issue: "Essentially without exception in eukaryotic systems, the RBE of high-LET radiation is a strong function of dose, increasing as the dose decreases." NCRP 64 (1980). Our analyses of Shellabarger et al (1969) and Montour et al (1977) studies are also described in more detail below in section 7.10.2 (see Figs. 30 to 33) as well as in Annex III, part 6.

In the BEIR III (1980) report the neutron RBE for the L-L model of leukemia incidence is also estimated by the ratio of two parameters in a linear dose-response model: $\hat{\theta} = \hat{\beta}_2/\hat{\beta}_1$ where β_1 is the coefficient of the gamma dose, D_γ, and β_2 is the coefficient of the neutron dose, D_n. However, since the precision of the estimates of the numerator are only modest ($\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$ $^{-}$ 3.5, j = 1,2) the point and interval estimates of θ are obtained by weighted jackknife methods as well as by Fieller's theorem. The point estimate given in BEIR III is $\hat{\theta}$ = 11.3 (eq. V-11). The weighted jackknife estimate of θ is $\hat{\theta}$ = 9.42, i.e., the BEIR III estimate is inflated by about 20%. The weighted jackknife 0.95 CL are (-3.81, 22.65); Fieller's theorem gives a point estimate $\hat{\theta}$ = 11.30 and 0.95 CL of (2.82, 40.33). Note that no interval estimates of either the RBE or the DREF are presented in either the BEIR III (1980) report or the NIH Radioepidemological Tables (1985).

It is appropriate here to discuss one of the more curious sections of the BEIR III report (1980), namely the dose-response model of the LSS data on breast cancer incidence in women that is presented therein. The BEIR III Report offers the L-L model of breast cancer incidence as the, "model of choice": "Breast-cancer data offer little support for a dose-response model with strong upward curvature in D_γ. The dose-response curves for mammary tumors in female rats given total-body x and gamma irradiation [sic] lead to be linear [N.B. Shellabarger et al (1969) is cited]. Functions of D_γ and D_n fitted to the breast cancer incidence rates for Hiroshima and Nagasaki, standardized to the age distribution of both cities, suggested a relationship linear in both D_γ and D_n ... Linear model coefficients for D_γ and Dn did not differ significantly ..." (BEIR III, 1980). These inferences would seem to be instances of what Bacon has referred to as, "... illicit generalizations and hasty conclusions", or, in Whewell's phrase, an "... imposing delusion of received theory" (see section 2.1). However, the estimate of the coefficient of the neutron dose, D_n, for both the LQ-L and L-L models of breast cancer incidence is much less than its standard error: $\beta_3/\sqrt{Var(\beta_3)} = 0.63$, for the L-L model. As was the case with the LSS data (T65D dosimetry) on leukemia incidence and non-leukemia cancer mortality, the LQ-L model of the LSS breast cancer data overfit the data with respect to the rival L-L model (Tables V-8 and V-9 in the BEIR III report show that the decrements in the chi-squared statistics of the respective pairs of rival models are not significant). For the LQ-L models of non-leukemia cancer mortality and breast cancer the coefficients of D_γ^2 are negative. See Herbert, 1986c.

As remarked above the covariance matrices, $Var(\bar{\beta})$, for each of the BEIR III models of dose response of the LSS data for leukemia incidence (Table V-8), non-leukemia cancer mortality (Table V-9) and breast cancer incidence (Land et al, 1980) are incorrectly estimated: Each is multiplied by a heterogeneity factor, $h = \chi^2/(n-k)$. For most of these models, h $^{-}$ 1 so the error is not so important, however for the L-L model of breast cancer incidence, h = 0.525, so that the reported precision is spuriously high, $\hat{\beta}_3/\sqrt{Var(\hat{\beta}_3)} = 0.87$, instead of the correct value of 0.63. (We shall examine the Q-L model of the LSS breast cancer data in section 7.10.2 below.)

Thus, while it is true, as remarked in the BEIR III Report, that for the L-L model the estimated coefficient for the neutron dose does not differ significantly from the estimated coefficient for the gamma dose, implying that $\beta_3 = \beta_1$, (which is consistent with an RBE = 1) it seems important to note also that the estimated coefficient for the neutron dose does not differ significantly from zero - although this was not remarked in the BEIR III Report. Note that the sample evidence that $\beta_3 = 0$ is inconsistent with the a priori, or non-sample, evidence (T65D dosimetry) that $\beta_3 > 0$. That the coefficient of D_n is (much) less than its standard error suggests the presence of large random and/or systematic errors in the measurement of neutron dose in the T65D dosimetry (The DS86 dosimetry suggests that there was a systematic error in the T65D measurements of neutron dose in the LSS sample). As the coefficient of D_n is less than its standard error, the usual model selection criteria of regression analysis suggests that the term in D_n should be deleted from the model since the increase in bias in the estimates, $\bar{\underline{\beta}}$, $\hat{\underline{y}}$, etc., is less than the concomitant decrease in variance which results from deleting D_n. Thus, the mean squared error of estimate is decreased by omitting D_n (Montgomery & Peck, 1982).

A novel alternative is to retain the term in D_n by "pooling" it with that in D_γ to form the pseudo-dose, $D^* = (D_\gamma + D_n)$. (N.B. For the linear model of the LSS breast cancer incidence data in which the "dose" is D^* the chi-squared goodness-of-fit statistic lies in the extreme lower tail: $\chi_c^2 = 8.44$, df = 17, $P(\chi^2 < \chi_c^2|17) = 0.04$. But it is the practice to reject models for which $P(\chi^2 < \chi_c^2|df) < 0.05$ on the (quite reasonable) grounds that the data "fit" the model too well to have arisen by random sampling from the population described by the model: "R. A. Fisher has emphasized that a small value of χ^2, e.g., $\chi^2 < \chi^2_{0.05}$, should lead to the rejection of the hypothesis, the agreement being too good to be true. Too good an agreement may be due to the fact that the hypothetical distribution includes too large a number of parameters so that it should be possible to use a smaller number of parameters and still obtain an adequate description of the observed distribution." A. Hald, 1952. See also Finney (1971); Jeffreys (1957).) This procedure is equivalent to the conclusion that $\beta_3/\beta_1 = 1$ as noted in the BEIR III Report. However, such an inference is inadmissible unless the ratios $\bar{\beta}_j/\sqrt{Var(\bar{\beta}_j)}$ for both variables are well in excess of unity, i.e., unless both differ significantly from zero, and seems wholly contrary to any reported practice in regression analysis. (But it is perhaps consistent with the practice of "pooling" information on leukemia incidence with that on non-leukemia cancer mortality that was required in order to obtain stable $(\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)} > 2)$ estimates of the parameter vector $\underline{\beta}$ of the LQ-L model of non-leukemia cancer mortality, still another "salvage" maneuver adopted in the BEIR III Report (vide infra).) However, this procedure also over-states the true precision with which the effect of D_n on the incidence rate, say m, represented by the weight, β_3, can be estimated from the LSS data. This precision is correctly estimated by the ratio $\hat{\beta}_3/\sqrt{Var(\hat{\beta}_3)} = 0.63$. If it is thought necessary (say, on the basis of prior information) to retain D_n in the model, then it must be weighted by the "non-significant" coefficient estimate, $\bar{\beta}_3$, provided by LSS data or else a quasi-Bayesian procedure must be adopted, such as Mixed estimation, by which the weak data may be "strengthened" by non-sample information - or conjecture. (For instance, the conjecture, H_0: $\beta_1 \equiv \beta_3$, can be represented as the constraint, $\underline{r} = R\underline{\beta}$, on the L-L model where $\underline{r} = 0$, $R = (0, 1, -1, 0)$, $\psi \equiv [0]$. The conjecture, H_0: $\beta_3 \equiv 0$, can be represented as the constraint, $\underline{r} = R\underline{\beta}$, where $\underline{r} = 0$, $R = (0, 0, 1, 0)$, $\psi \equiv [0]$. This gives the rival models L-L, L^* and L, respectively. The Bayesian procedure makes quite clear that the relation $\beta_1 \equiv \beta_3$ is derived more from non-sample information rather than from sample information, as is implied in the BEIR III report.

If the neutron RBE for the L-L model of breast cancer incidence is estimated by the same procedure as was done for the L-L model of leukemia incidence in the BEIR III Report, the point estimate is $\hat{\theta} = 1.42$. The point and interval estimates obtained by the weighted jackknife method are $\hat{\theta}_J = 1.08$ with standard error, $\sqrt{Var(\hat{\theta}_J)} = 2.01$. Obviously, there is little empirical support for the estimates of neutron RBE = 1 for breast cancer from the BEIR III LSS data. (It should be remarked that a correct interpretation of the information on the coefficient of the neutron dose, D_n, in the L-L model of the LSS data on breast cancer would have provided one of the earliest clues that the level of neutron dose may have been over-estimated in the T65D dosimetry.)

7.5.3.3 Why does the form 'y = a + bx' occur so often in radiobiological models?

"Few asked themselves what the long term future might be for a method which - to be brutal - solves the wrong equation. 'Give me an answer!' is the demand. So the linear theory obliges, hoping that no one will notice when it's the wrong answer."

I. Stewart, 1989

"If the only tool you have is a hammer, you tend to treat everything as if it were a nail."

A. Maslow, 1966

As remarked above, it appears from our review of the studies listed in Table 1 that low-dose stochastic radiation effects are commonly described by dose-response models and high-dose stochastic radiation effects by isoeffect models. However, it appears that for either dose-response or iso-effect models in received practice most investigators seek to summarize the observations by a conceptually simple, computationally tractable, linear regression model of the form y = a + bx. (Note that we have omitted representation of the random part of the response, y, since it is effectively ignored in most of received practice, although as we have remarked above, it is the nature of the distribution of the random part of the response that determines the form of the link function of the appropriate generalized linear model, e.g., Normal, binomial, Poisson, etc. See, for instance, McCullagh and Nelder, 1989.)

The above quotation from the English mathematician, Ian Stewart, adverts to the historical fact that the always abundant empirical evidence that Nature is intrinsically - and exuberantly - non-linear has been, until quite recently, meticulously disregarded by nearly all scientists. Stewart further notes that, "Classical mathematics concentrated on linear equations for a sound pragmatic reason: it couldn't solve anything else ... So docile are linear equations that the classical mathematicians were willing to compromise their physics to get them. So classical theory deals with shallow waves, low-amplitude vibrations, small temperature gradients." However, the recent advent of digital computers that can implement the computations - and graphics - required for the solution of the refractory non-linear equations, has enabled the scientists (physicists in particular) to advance into fields of enquiry that were hitherto foreclosed; namely, non-linear dynamics and non-equilibrium thermodynamics, and to study such (classically) counter-intuitive phenomena as deterministic chaos, dissipative structures, solitons, and fractal geometry.

But this is exactly the situation we have described in this report: It seems that more than a few in the radiobiology community have often ignored the abundant evidence that the advent of the digital computer has made it possible to readily supersede the ingenuous linear models, "y = a + bx" - the "conceptual box" provided by their professional educations - into which a wide variety of radiation response data has hitherto been rudely thrust (by "strenuous and devoted" attempts) with the appropriate generalized linear (probit, Poisson, etc.) and non-linear models. These radiobiologists were often willing to compromise their facts to get the more "docile" linear equations. As Hegel might have remarked, "So much the worse for the facts." (N.B.: It must be noted that the non-linear models of non-linear dynamics are non-linear in the (predictor) variables, e.g., x^2, sin x, etc., whereas the non-linear regression models of radiation biology are (usually) non-linear in the parameters. The generalized linear models are linear in the parameters but the response variable may be a non-linear function (e.g., probit) of the observed response.)

We have already examined several unfortunate consequences of this feature of the received practice in the cases of the Shellabarger et al (1969) and Montour et al (1977) models of radiation-induced mammary neoplasia in female Sprague-Dawley rats. This is a low-dose effect in which the observed response has a Binomial distribution (see Figs. 13a and 13b). We must now re-direct the reader's attention to the respective received models of the LQ hypothesis for the high-dose effects of fractionated radiation treatments in which the observed response has a Binomial distribution; we must examine the isoeffect curve of the Fe-plot for the multifraction (N $\geq$ 1) LQ hypothesis:

$$D^{-1} = (\alpha/\ln S) + (\beta/\ln S)(D/N).$$

For the low-dose effects the observed responses have a Poisson distribution and we examine the dose-response curve of the Chapman plot for the single-fraction (N = 1) hypothesis:

$D^{-1}\ln S = \alpha + \beta D$.

In the F_e-plot the variable S is the unknown (and unknowable) level of survival at dose D of an unidentified cell-population. In the Chapman plots the levels of S are well-known as is also the identity of the cell-population. However, both models have achieved the simple form, a + bx, for the right-hand side - the linear predictor. For both, the received practice is to obtain sample estimates of the "slope" and "intercept" terms either by least squares or by graphical methods (see Figs. 2a and 2b, respectively). However, the Chapman model implies (see Neter and Wasserman, 1974) the a priori - and arbitrary - assumption that the n observations have weights $w_1 = 10^6$, $w_i = D_i^{-2}$, $2 \leq i \leq n$, although there is no empirical warrant for such an assumption. And the equation of the F_e-plot implies that <u>only</u> the ratio, α/β, of the intercept and slope can be estimated - since lnS is unknown (and often unknowable).

Note that for Binomial-distributed responses for $N > 1$ the received data-analytic practice in radiation biology "solves" the statistical - and computational - problems of estimation and inference that reside in the multivariate deterministic part, μ_i, <u>and</u> a non-Normal distribution of the random part, ϵ_i of a response, $y_i = \mu_i + \epsilon_i$, by simply "transforming", or "collapsing", the <u>dose-response model</u> that describes the variation in the observed response to the changes in the dose, D, and other covariates, etc., prescribed in the experimental design <u>and which is the central matter at issue</u>, to an <u>isoeffect model</u> that describes the variation in the dose, $D(\pi)$, which is required to achieve a fixed level of response, say π, as the covariate, N (and T), changes - and which is a derivative issue. This "collapsing" maneuver effectively <u>discards much of the data</u> and, of course, also the information on dose-response that is contained therein - suggesting (again) that data may not be taken too seriously by some members of the radiobiology peer-group. We shall return to this issue below in section 12.

The Cell Inactivation Plot of the LQ model of cell-survival does, of course, describe the variation in response with dose at a fixed level of fractionation: $N = 1$. That is, it addresses the central matter at issue: the construction of a model of radiation dose-response. However, the maneuvers by which this method obtains the parameter estimates for the LQ model implies - entails, actually - two questionable assumptions whose validity is not - and <u>cannot</u> - be demonstrated. That is, the ordinary least square estimates of the parameters β_1 (=α) and β_2(=β) in the equation of the Cell Inactivation Plot that are obtained implies that the sample observations have the set of <u>a priori</u> weights. 1) i=1, $w_1 \longrightarrow \infty$, since the observed responses, m_i, $1 \leq i \leq n$, are first transformed to survivals, S_i, by imposing the <u>constraint</u>, $m_1 = e^{\beta}0$; and 2) $w_i = D_i^{-2}$, $2 \leq i \leq n$. Neither entailment would appear to carry any empirical warrant and in Annex IV, part 6, we have demonstrated that they do not. Thus, it seems to be the case that the current praxis in dose-response modelling by the radiobiology peer-group may often require rather extensive transmogrifications of the raw data in order to achieve a form that is consistent with the simple equation, $y = a + bx$, in which the estimates of the slope, b, and intercept, a, can be readily obtained either by ordinary (unweighted) least squares or by graphical methods. (To quote Acton (1959) again, 'Some data fit lines naturally, but others have lines thrust upon them.") But although these estimates are easy to get, they are hard to swallow. (Good models are so rare because bad models are so easy.)

Some additional support for our conjecture can be found in the early advocacy of the LQ model in which Fowler (1984) remarks that, "The dose-response formulae that could be used with <u>statistical legitimacy</u> [emphasis added] include (1) the linear quadratic model ... [and] For present purposes the LQ model offers (a) valid approximations for all doses likely to be used in radiotherapy, (b) only two parameters to be determined, and (c) considerable convenience in practical applications." But of course the current literature on the LQ model offers no acceptable empirical evidence for the statement (a). On the contrary, our recent review of that literature discloses that the goodness-of-fit of an LQ model has rarely been assessed by statistically adequate measures. And in those studies of the LQ model that we have re-analyzed (see Annex II, parts 3, 4, and 5), in which we carefully examined, by statistically adequate measures, the issue of concordance, we found that the LQ model doesn't "fit" - either at all in some cases, or as well as a rival model in others.

The putative attractive features of the LQ model cited in statements (b) and (c) seem to support the above conjectures concerning deployment of the model y = a + bx. First, as we have noted, the fact that the received form of the LQ model of cell survival (N=1) has only two, rather than three, parameters is a consequence of the received practice of transforming the survival data by the "survival transformation", $m_i/m_1 \longrightarrow S_i$, $1 \leq i \leq n$, which is equivalent to the imposition of the a priori and arbitrary constraint $\beta_0 = \log m_1$, which, as we have remarked, degrades the "fit" (inflates RSS) of the model to the data (see section 7.2.2 and Annex IV, part 6). Second, the fact that the multifraction ($N \geq 1$) LQ model has only two parameters is a consequence of omitting the time factor which suggests that the model is likely to be misspecified - it underfits the data. (It will be recalled that underfitting entails biased estimates of $\underline{\beta}$; overfitting entails inflated variances, Var($\hat{\underline{\beta}}$) and Var($\underline{x}_i^T\hat{\underline{\beta}}$) and "... both improper omission and indiscriminate inclusion of variables in a model can lead to compromised inferences." Robins and Greenland, 1986) Third, the usual desiderata that guide the selection of variables in statistically adequate modelling practice, parsimony (which according to Jeffreys (1960) identifies the "correct" model), and reduced variance in the sample estimates $\hat{\underline{\beta}}$ and $\underline{x}^T\hat{\underline{\beta}}$ of the parameters and response, respectively, are never cited as arguments for the advantages of, "only two parameters to be determined". Thus, it would seem that the decisive advantage offered by the LQ model that is perceived by its advocates is described in statement (c) and therefore that the "considerable convenience in application" refers to the fact that the received version of the LQ model has the form y = a + bx, as remarked above. This conclusion is reinforced by a remark in a subsequent review by Fowler: "There are now good time factors available for tumors and acute reactions but there is a price to pay. Four parameters have to be specified instead of one (see below)" (Fowler, 1989). However, no discussion of the "price" in the only terms that are relevant - inflation of the variance (if the model is overfit) and inflation of the bias (if it is underfit) - is presented.

Kuhn (1962) has described so-called Normal Science as, "... a strenuous and devoted attempt to force nature into the conceptual boxes provided by professional education." The perspicuity of his view of current scientific practice can be seen quite clearly in the current literature on regression models of radiation dose-response in which, in study after study, the data had "straightness thrust upon them" suggesting that the professional education of most investigators has included, as statistical methodology, just a box marked "y = a + bx", in which the response y has a (conditional) Normal distribution and the sample estimates of the parameters a and b are obtained by ordinary (unweighted) least squares methods. But (to mix a metaphor), this particular "box" is a Procrustean bed. It does not - and cannot - span the problems that confront most investigators when they come to the construction and evaluation of the models that they have selected to convey the information contained in the data obtained from the dose-response experiments which they (or others) have performed in which the response has a non-Normal (Binomial or Poisson) distribution and depends upon more than one predictor variable.

On historical grounds alone this inadequacy should come as no surprise. After all, the ordinary least squares methodology is nearly two hundred years old. (Legendre first proposed estimating the slope parameter β by minimizing the sum of squared residuals, $\underline{e}^T\underline{e} = \Sigma e_i^2$, over values of β, in 1805.) Like those of other scientific and technological artifacts that also originated in the nineteenth century - for example, the X-ray machine - the subsequent developments in data analysis have improved quite markedly on the original methods of "naive least squares", in effectiveness, as well as in subtlety and sophistication, in the intervening generations, in step with the corresponding increases in the subtlety and sophistication of our understanding of the world and how to manipulate it. Thus, the modern regression model is multivariate: $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, where $\underline{y}$ and $\underline{\varepsilon}$ are (n*1) matrices, X is an (n*k) matrix and $\underline{\beta}$ is a (k*1) parameter matrix as described above, and more sophisticated goodness-of-fit measures, diagnostics, etc., as well as the parameter estimates are required for a full understanding of the information in the data that is - or is not - conveyed by the model.

On the other hand, as remarked above, the ubiquitous presence in the literature of variations on the form "y = a+bx" can, perhaps, be taken as evidence of a profound (even when subconscious) desire on the part of the investigators for (to again borrow Bunge's locution) "... a single syntactic form that is invariant under a wide variety of semantic transformations." (Such a desire is, of course, a basic motivation of mathematical practice: "Mathematics is the art of giving the same name to different things" H. Poincaire.) But this desire for an invariant syntactic form is, of course, satisfied by the generalized linear model (McCullagh & Nelder, 1983), discussed above in part 4, Statistical methods. I. The generalized linear model (GLM) is a central theoretical concept - and methodological construct - in modern statistical modelling. For a sample of size n of stochastic responses, y_i, $1 \leq i \leq n$, we have $y_i = \mu_i + \epsilon_i$; $g(\mu_i) = \eta_i = \underline{x}_i^T\underline{\beta}$ where, as above, μ_i is the deterministic part and ϵ_i is the stochastic, or random, part of the response. $E(y_i) = \mu_i$ is the expected value of y_i, g(.) is the link function, with form determined by the form of the distribution of ϵ_i, e.g., if ϵ_i has a Binomial distribution g( ) is the probit transform $z_i = \Phi^{-1}(\pi_i)$. η is the linear predictor. In the usual matrix formulation the generalized linear model, $g(\mu) = \underline{x}_i^T\underline{\beta}$, $1 \leq i \leq n$, immediately redresses the two ontological weaknesses in current practice that were remarked above: 1) the inherent multivariate nature of the deterministic part of the response is correctly taken into account by the linear predictor, $\eta_i = \underline{x}_i^T\underline{\beta}$, $1 \leq i \leq n$, and 2) the inherent non-Normality of the distribution of the random part of the response is correctly accounted for by an appropriate selection of the form of the link function, g(.). To continue the original analogy with x-ray machinery, the GLM represents a cognitive MRI machine - a state-of-the-art epistemological tool.

7.6 Some further comments on goodness-of-fit. Residuals.

"The main questions which we ask of residuals ... are broadly:

(i) Does the fitted model adequately describe the data?

(ii) Are the data adequate for determining the model? Here we may be concerned with the need for data editing ... we may also be concerned about whether more observations are required for model discrimination.

(iii) Is the assumption about the nature of the error distribution correct and if not is a re-analysis necessary?"

P. J. Harrison, 1968

Let us now examine more closely the epistemological weaknesses - and their consequences - in the modelling praxis that is described in the radiobiological literature, weaknesses that have been disclosed to the secondary analyses presented in Annexes I-IV to the present report. These are, of course, the weaknesses in the methods by which we learn from experience; that is, in the methods by which we construct models that accurately and usefully epitomize that experience - the data and the circumstances under which it was obtained - and hence to that degree are "believable".

The greater part of the variation in the observed response will, of course, in the general case, be accounted for by any model - or the law that it articulates - that is reported in the literature, leaving always an unexplained balance. But, the "... ground for accepting and generalizing a law is ... a quantitative one: how much of the observed variation needs to be explained by the law before we can accept the law?" (Jeffreys, 1961). (As remarked above, investigators trained in the more deterministic sciences - such as physics - generally tend to believe that all of the observed variation can - or should be - explained deterministically. It often appears that they do not recognize and hence do not try to describe, or model, the random, or stochastic, part of the observed response - Jeffreys' "unexplained variation". (N.B.: But random - chaotic - behaviour of solutions of deterministic systems is now understood to be an inherent feature of many non-linear systems. That is, the occurrence of randomness in the behaviour of the system as it evolves in time does not require the presence of a stochastic forcing term; it resides in the non-linearity of the system itself. As we suggested in the discussions of publication bias above, the fact that most physicists lacked an interest in the more aleatory aspects of the behaviour of simple dynamical systems, delayed the "discovery" of deterministic chaos for at least seventy years.) An essential, and quite straight- forward, criterion for deciding whether either a hypothesis, or the model by which it is articulated, is "believable" - or useful - was offered just four hundred and fifty years ago in the preface to an exposition of the heliocentric hypothesis of Copernicus by Osiander (1540): "Nor is it, to be sure, necessary that these hypotheses be true, or even probable; but this one thing suffices, namely, whether the calculations show agreement with the observations." (Actually, Democritus seems to have first established the methodological principle that a deductive theory or explanation must, "save the phenomena", that is, must be in agreement with experience. Thus, goodness-of-fit has been an epistemological - as well as statistical - issue for at least the past two and a half millennia.)

The statistician K. Pearson, an instrumentalist like Osiander (and Mach), echoed the Osiander criterion in his remarks in the Grammar of Science (1892/1911): "The only final test we have of the truth of any law, of the sufficiency of its description, the only proof ... is the actual comparison of the results of the formula with the facts themselves." (N.B.: Sir Harold Jeffreys, the great geophysicist, has remarked (1957) that the Grammar of Science was, "the outstanding work on the subject [of scientific induction]." And it is still well worth reading.) Pearson, of course, also developed the so-called Pearson chi-squared goodness-of-fit test that implements this criterion for a wide variety of data:

$$\chi^2 = \Sigma(O_i - E_i)^2/E_i$$

where χ^2 is the so-called Pearson chi-squared statistic. It is distributed as a gamma variate (Hastings and Peacock, 1975) for which the mean is ν and the variance is 2ν, where ν is the number of degrees of freedom, O_i is the observed, and E_i is the expected, number in the ith category, $1 \leq i \leq n$.

For regression models we have the generalized Pearson chi-squared statistic, a sum of squared residuals:

$RSS = \chi^2 = \Sigma(y_i - \hat{\mu}_i)^2/Var(\hat{\mu})$ (See Chatterjee and Hadi, 1988.)

For a Normal theory model (Chatterjee and Hadi, 1988), the generalized Pearson statistic is

$$\chi^2 = \Sigma(y_i - \hat{\mu}_i)^2/Var(\hat{\mu}_i) = \Sigma(y_i - \hat{\mu}_i)^2/h_i\sigma^2$$

where σ^2 is the variance and h_i is the i^{th} hat matrix diagonal. The respective forms of χ^2 for models in which y_i has a non-Normal distribution are similar to that for the Normal theory model, namely

$$\chi_i = (y_i - \hat{y}_i)/\sqrt{Var(\hat{y}_i)}$$

For a dose-response model of a Binomial response, we have

$$\chi^2 = \Sigma(r_i - n_i\hat{\pi}_i)^2/n_i\hat{\pi}_i(1-\hat{\pi}_i)$$

where r_i is the observed number of responders in n_i at risk at $\underline{x}_i T$ and $\hat{\pi}$ is the estimated proportion $(0 \leq \hat{\pi}_i \leq 1)$ of responders expected at $\underline{x}_i^T$ - if the model is correct.

For a dose-response model of a Poisson response

$$\chi^2 = \Sigma(y_i - c_i\hat{m}_i)^2/c_i\hat{m}_i$$

where y_i is the observed number of responders at $\underline{x}_i^T$, c_i is the "concentration" and $\hat{m}_i$ is the estimated response rate at $\underline{x}_i^T$ - if the model is correct.

The elements of χ^2, namely, $\chi_i = (r_i - n_i\hat{\pi}_i)/\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$ and $\chi_i = (y_i - c_i\hat{m}_i)/\sqrt{c_i\hat{m}_i}$ are the more familiar Pearson chi-residuals, that is, $\chi^2 = \Sigma\chi_i^2$. Note that the sums, $\Sigma\chi_i^2$, Σd_i^2 and Σg_i^2, where d_i and g_i are the deviance and Freeman-Tukey residuals, respectively, described in Table 3 are, for models that "fit" the data, each distributed as χ^2 on (n-k) degrees of freedom where n is the number of levels of $\underline{x}_i^T$ $(1 \leq i \leq n)$ and k is the number of estimated parameters in the model. N.B. For the chi-squared distribution on ν degrees of freedom the expected value (mean) and variance are ν and 2ν, respectively.

In the sequel we will represent residuals for non-Normal theory models by $e_i^* = \chi_i$, etc. See Tables 3a and 3b and Annex II, parts 3 and 5. N.B. For parametric models of dose-response, the observed and estimated levels of response are usually compared in terms of the differences, say $e_i = (y_i - \hat{\mu}_i)$ - and functions thereof. However, for parametric models of time-to-failure the comparison is often in terms of ratios, say $e_i^{\#} = y_i/\hat{\mu}_i$ - and functions thereof. For example, for a two-parameter Weibull model, $e_i^{\#} = [t_i/\exp(\underline{x}_i^T\hat{\beta}]^{\delta}$, for an uncensored (exact failure) and $e_i^{\#} =$

Table 3a. Case Statistics of Fit. Residuals. Goodness-of-Fit Criterion. Regression Diagnostics.

Form of distribution of Response	Type of Residual#	Form of Residual(11,19,27,44,57,80)	Residual Plots
1. Normal $-\infty < y_i < \infty$	Deviance Pearson Standardized	$e_i = (y_i - \hat{\mu}_i)$ $e_i^* = e_i/\sigma\sqrt{(1-h_i)}$	e_i vs $\hat{\mu}$ y_i i x_{ji} z^i
2. Poisson $0 \leq y_i < \infty$	Deviance Pearson Freeman-Tukey##	$d_i = \mathrm{sgn}(y_i-\hat{\mu}_i)2[y_i\log(y_i/\hat{\mu}_i) - (y_i-\hat{\mu}_i)]^{1/2}$ $\chi_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}$ $g_i = \sqrt{y_i} + \sqrt{(y_i+1)} - \sqrt{(4\hat{\mu}_i+1)}$	d_i, χ_i, g_i vs $\hat{\mu}_i$ y_i x_{ji} i z_i
3. Binomial $0 \leq \pi_i \leq 1.0$ $0 \leq r_i \leq n_i$	Deviance Pearson###	$d_i = 2\sqrt{[r_i\log(r_i/n_i\hat{\pi}_i) + (n_i-r_i)\log[(n_i-r_i)/(n_i(1-\hat{\pi}_i))]}$ $\chi_i = (r_i-n_i\hat{\pi}_i)/\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$	d_i, χ_i vs $\hat{\pi}_i$ i x_{ji}

# "... residuals are mainly used to look for what might be going on beyond what is already in the model. ... The important thing is to look at the residuals, details of definition matter much less" (J. Tukey, 1968). Or, "... in all human undertakings the issue is decided in the details" (H. Poincaire, 1892).

## See Velleman and Hoaglin (1981)

### "The simplest, most direct and probably most useful definition of residual in the context of binomial or binomial-like data is the Pearson residual ..." (McCullagh and Nelder, 1983)

Table 3b. Aggregate Statistics of Fit. Residual Sum of Squares (Sampling Measure). Goodness-of-fit Criterion.

Form of Distribution of Response	Type of Measure	Form of Measure	Sample Distribution of Measure
1. Normal $-\infty < y_i < \infty$	Sum of Squared Residuals	$RSS = \Sigma e_i^2$	$RSS/\sigma^2 \sim \chi^2(n-k)$
2. Poisson $0 \le y_i < \infty$	Sum of Squared Residuals	$RSS = \begin{matrix} \Sigma d_i^2 = D \\ \Sigma \chi_i^2 = \chi^2 \\ \Sigma g_i^2 = F^{\#} \end{matrix}$	$RSS \sim \chi^2(n-k)^{\#\#}$
3. Binomial	Sum of Squared Residuals	$RSS = \begin{matrix} \Sigma d_i^2 = D \\ \Sigma \chi_i^2 = \chi^2 \end{matrix}$	$RSS \sim \chi^2(n-k)^{\#\#\#}$

# N.B.: F denotes "Freeman-Tukey", not the usual F, or variance-ratio, statistic.

## N.B.: For "sparse" Poisson data (many cells with $y_i = 0$ or 1) the degrees of freedom for the chi-squared approximation to the sampling distribution of Σg_i^2 are reduced by the Tukey correction factor, $c = \Sigma^*(1 - \hat{\mu}_i)^2$, where the summation $\Sigma(1 - \rho_i)^2$ is over all cells for which $\hat{\mu}_i < 1$. (See Velleman and Hoaglin, 1981.)

### N.B.: For "sparse" Binomial data (all cells with $y_i = 0$ or 1), the Hosmer-Lemeshow statistic should be calculated.It has a chi-squared distribution on (g-2) df where $g \le 10$ is the number of groups of observations. (See Hosmer and Lemeshow, 1989. See also Finney, 1971.)

$= [t_i/\exp(x_i^T\hat{\beta})]^{\hat{\delta}} + 1$ for a right-censored observation. The $e_i^{\#}$ are often referred to as generalized residuals (See Cox and Snell, 1968; Lawless, 1982). The $e_i^{\#}$ are independently Exponentially distributed with unit mean. Recall that for the Normal theory model discussed above the residuals $e_i = (y_i - \hat{\mu}_i)$ are distributed Normally with zero mean.

7.7 Model discrimination I. The Akaike Information Criterion

For many data sets there are two or more rival models, all of them often equally plausible, a priori, and it is necessary to discriminate between them empirically on the evidence of both sample and non-sample information.

Rival models can be classified as either nested or non-nested. Two rival nested models are the Poisson loglinear models, $m_i = \exp(\beta_0 + \beta_1 D_i)$ and $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$. Two rival non-nested models are the Poisson loglinear model, $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$, and the Poisson non-linear model, $m_i = \beta_0\{1-[-\exp(\beta_1 D_i)]^{\beta}2\}$. In all four models m_i is the Poisson rate parameter.

For two nested rival linear models, for which the respective linear predictors differ by one term, the sampling distribution of the difference, ΔRSS, in the respective values of RSS is adequately approximated by the Pearson chi-squared distribution on 1 df, for which the 0.95th quantile is 3.84. For any two nested models, the criterion can be generalized in obvious ways. Thus RSS can be used as a discrimination criterion in model selection.

However, for two non-nested models, the decrement, ΔRSS, in the sum of squared residuals, $e_i^{*2}(\chi_i^2, d_i^2, \text{ or } g_i^2)$, cannot be used to discriminate between them. The Akaike Information Criterion, AIC, is a useful discriminant for non-nested models. But, it is a criterion of the quality of the model; it does not provide a formal test of goodness of fit (Akaike, 1974, 1977, 1985; Gilchrist, 1984). That is, the usefulness of the AIC criterion does not depend on the sampling distribution of a statistic such as χ^2, as does the decrement ΔRSS. The AIC is derived from information theory and will identify that one of several rival models for which the mean squared error of prediction is least. It will be recalled that the mean squared error (MSE) of an estimate, say $\hat{y}_i$, is

$$MSE = Var(\hat{y}_i) + Bias^2(\hat{y}_i)$$

The variance of the estimate is inflated for models that overfit the data; the bias of the estimate is inflated for models that underfit.

The AIC is defined as

$$AIC = -2\ln L(\hat{\beta}) + 2k = -2\ln L + 2k$$

where $L(\hat{\beta})$ is the maximum likelihood function for the model and k is the number of parameters. The AIC may be written in terms of the sum of squares of the deviance residuals:

$$AIC = D(\hat{\beta}) + 2k = \Sigma d_i^2 + 2k$$

where $D(\hat{\beta})$ is the deviance and d_i is the deviance residual.

The AIC identifies the "best" model as that one for which AIC is a minimum. Akaike (1977) has pointed out that when the concordance of the maximum likelihood estimates of β is measured by the Pearson chi-squared statistic, χ^2 (=RSS) then AIC = χ^2 - 2(n-k) = RSS - 2(n-k). That is, AIC is a joint measure of the concordance of the model and data and a factor which describes the relation between the size of the sample (n) and the dimension of the model (k).

Note that the AIC criterion combines the joint desiderata of accuracy and parsimony in a model. Thus, the AIC implements Ockham's Razor. The first factor, -2logL, describes the fidelity of the model to the sample observations. The second factor, 2k, describes the degree of parsimony in the model. That is, if two models fit equally well, the model with the fewer parameters, k, is chosen - on the basis of the Jeffrey-Wrinch simplicity postulate: "... the simplest law is chosen because it is the most likely to give correct predictions; ... the choice is based on a reasonable degree of belief" (Jeffreys, 1961). The choice of the more parsimonious model may be also based on the fact that the variance of the predicted response increases with k (Montgomery and Peck, 1982). Efron (1986) has shown that using AIC to discriminate between rival models amounts to selecting the model that gives the minimum mean square error of prediction (over the sample). Several forms for AIC are given in Table 4a.

Table 4a. Aggregate Statistics of Fit. Information Theory Measure (Non-sampling Measure).

Form of Distribution of Response	Name of Measure	Form of Measure
Normal	Akaike Information Criterion (AIC)[#]	$AIC = D + 2k$
Poisson		$AIC = D + 2k$ $AIC = \chi^2 - 2(n-k)$
Binomial		$AIC = D + 2k$ $AIC = \chi^2 - 2(n-k)$

[#] Model discrimination criterion (Nested and Non-nested rivals). Minimum AIC (MAICE) selects the model for which the mean squared error of prediction (in the original data) is least. (Akaike 1974, Efron, 1986)

N.B. It has been shown that the cross-validation and Akaike criteria are asymptotically equivalent. (Stone, 1976)

Table 4b. Aggregate Statistics of Fit. Predicted Residual Sum of Squares (Non-Sampling Measure).

Form of Distribution of Response	Name of Measure[#]	Form of Measure
1. Normal	PRESS	$Press = \Sigma e_{(i)}^2 = \Sigma e_i^2/(1-h_i)^2$
2. Poisson	PRESS	$PRESS = \Sigma \chi_{(i)}^2 = \Sigma \chi_i^2/(1-h_i)$
3. Binomial	PRESS	$PRESS = \Sigma \chi_{(i)}^2 = \Sigma \chi_i^2/(1-h_i)$

[#] Model discrimination criterion (Nested and Non-nested rivals). Minimum PRESS selects the model for which the degradation of predictive performance in new data ("regression of regression") is least. The procedure is sometimes refereed to as jackknife validation. (Gilchrist, 1984)

Note that although the AIC is a discriminator between linear and non-linear models - i.e., non-nested rivals - it does not address the question of the optimal functional form of the model, e.g., linear vs non-linear models. (Akaike, 1974; Akaike, 1985) Further discussion and examples of the use of the AIC criterion in model discrimination are presented in Annex II, parts 3 and 5 for models of Binomial and Poisson responses, respectively.

7.8 Model discrimination. II. Bayesian methods in dose-response modelling.

"In a fundamental sense the Bayesian procedure for changing initial beliefs is a learning model of great value in accomplishing a major objective of science - learning from experience."

A. Zellner, 1971

In the classical approach to model selection, as outlined first in part 4, the selection procedure is cast in the form of a hypothesis test in which it is necessary to discriminate on the basis of a set of observations, between the null hypothesis, H_0, and the alternative hypothesis, H_1, with probabilities of Type I and Type II errors specified a priori. These are the hypotheses H_0: model M_0 and the data [y, X] are not inconsistent and H_1: M_0 and [y, X] are inconsistent.

However, instead of considering the prior probabilities of errors incurred in taking a decision on the null hypothesis, H_0, we may wish to consider the probability of the hypothesis H_0 itself. Such a concept does not arise in classical sampling theory because of its requirement for a repetitive element in the definition of probability. (Can a theory be assigned a probability? Can the class of theories be closed and so be "... a suitable subject for probability statements." J. Nelder, 1986. N.B.: There are two current usages of the concept and term "probability". One is concerned with propositions that relate to outcomes of some kind of chance set up, e.g., a coin toss. The second is concerned with the measure of belief in a proposition. Shafer (1976) has named these aleatory and epistemic probabilities, respectively.) However, the concept arises quite naturally in Bayesian theory. (As we have seen in Jeffrey's simplicity postulate in section 3.3 above: "The set of all possible forms of scientific laws is finite or enumerable, and their initial probabilities form the terms of a convergent series of sum 1 ... the order of decreasing initial probabilities is that of increasing complexity.") Therefore, in this case we are led to consider the prior probability, $P(M_0)$ and the posterior probability $P(M_0|y)$, of the model M_0, given the observations [y, X]. This concept provides a useful method for comparing two rival models, say M_0 and M_1, of a given set of data, [y, X]. It is assumed that M_0 and M_1 are mutually exclusive and exhaustive: $P(M_0) + P(M_1) = 1.0$, and that $P(y) = P(y \& M_0) + P(y \& M_1)$. The estimates of the prior probabilities, $P(M_0)$ and $P(M_1)$ are obtained from prior experience - either experimental or non-experimental observations - or from relevant theory, or from introspection. $P(y|M_i)$, $i = 1,2$, is the average likelihood of y given M_i. Note that when the form and parameters of the model M_i are derived from experiments on target populations other than human, $P(M_i)$ includes the strength of prior belief about the validity of the inter-species extrapolation - the so-called "mouse to man" extrapolation of the model M_i.

Unfortunately, it is not infrequently the case that none of the several available rival models of a given set of observations can be rejected on the sample evidence alone. Therefore, in model discrimination it is most important to consider non-sample evidence as well as the sample evidence for the model that is summarized by the various measures of goodness of fit, for instance, the RSS. As Leamer (1978) has remarked: "... the data evidence should incrementally affect your opinion." (emphasis added). Thus, we wish to examine the evidence of the sample for the rival and received models in the context of the non-sample evidence for each.

Bayesian methods provide a formal means for bringing together conceptual evidence and empirical evidence on a given issue (See Zellner, 1971; Leamer, 1978; Gilchrist, 1984). For example, the posterior odds ratio $P(M_1|y)/P(M_0|y)$ for a model M_1 with respect to model M_0 combines the conceptual evidence for each of the two models, represented by the prior odds ratio $P(M_1)/P(M_0)$, with the sample evidence in the observation matrix, [y, X], for each of the rivals represented by the likelihood ratio, or Bayes factor, B:

$$P(M_1|y)/P(M_0|y) = [P(y|M_1)/P(y|M_0)] * P(M_1)/P(M_0)$$

where $B = P(\underline{y}|M_1)/P(\underline{y}|M_0)$. (If one is able, and willing, to specify the relative loss, say, l_{jk}, that may be incurred in the selection of model M_j, with parameter vector $\underline{\beta}_j$, as correct when it is M_k with parameter vector $\underline{\beta}_k$ that is correct, this information may be incorporated into the equation for the posterior odds ratio as a loss ratio.) The Bayes factor may be usefully regarded as a statistical artefact that enables the investigator to think more closely about the data in the context of the scientific background represented by the prior odds ratio. Note that the two rival models may be either nested or non-nested. The sample evidence favors M_1 if $B > 1$. This method of comparison forces a response to the most insistent question: How strong is the evidence of the sample [$\underline{y}$, X] for the model M_i, relative to the prior odds ratio?

The Bayes factor can be written as

$$B = (RSS_0/RSS_1)^{n/2}\ n^{(k_0-k_1)/2}$$

where RSS_j is the sum of squared residuals for the model with k_j parameters of sample size n; j = 0, 1. (See Leamer, 1978.) See also Annex II.

Just as in the case of the AIC, the sample evidence for the model can be described by a joint measure of the concordance, of the data [$\underline{y}$, X] and model, M_i, the residual sum of squares, RSS_i, and a factor which describes the relation of the dimension of the model, k_i, to the size of the sample, n. It is of interest to note that the difference in the respective values of AIC_i for two models, M_i, i = 0,1, has a form rather similar to the Bayes Factor, B:

$$\Delta AIC = \ln[(L_0/L_1)^2 e^{(k_0-k_1)}]$$

We do not include in the body of the report any examples of the use of these Bayesian methods. However, further discussion and examples of the use of the posterior odds ratio in model discrimination can be found in Annex II, parts 3 and 5, for models of Binomial and Poisson responses, respectively.

Still another method of discrimination between non-nested rival models is the method of embedding (Gilchrist, 1984). An example of the use of this method in discriminating between rival models of the random part of a response was given in section 7.1 (see Fig. 6c). A proposed use of the method in discriminating between rival models of the deterministic part of a Poisson response (LQ vs Target theory models of cell-survival) is described in section 14.2. See also Muirhead and Darby, 1987 and Preston, 1989.

7.9 Some further comments on the "believability thing" together with several illustrations.

"The fitting of laws to the data is still often done graphically, and this introduces unknown personal errors, while estimates of uncertainty, if any are given, are usually based on 'judgment', that is, guess-work."

Sir Harold Jeffreys, 1957

Table 2 presented the set of five criteria that must be satisfied by an adequate - that is, a believable - scientific model, which were described in part 3.2 - together with the respective statistical measures by which these criteria may be articulated and implemented.

In the present context the question that must be asked before any deployment of a model, is "Are the published point estimates of the model parameters, $\underline{\beta}$, point estimates of linear functions of model parameters such as the predicted levels of response, $\underline{x}_i^T\underline{\beta}$, and point estimates of non-linear functions of model parameters such as the ratio, β_1/β_2, 'believable' if it is not (or cannot be, vide infra) established that the regression model from which they are derived is concordant with the sample of observations on which the predictions and estimates are reported to be based and is consistent with a priori information on the process that generated them?" A review of the current literature on dose-response models discloses that the concordance, or goodness-of-fit, and consistency of a model are, quite surprisingly in view of the gravity of the clinical and public health enterprises which they inform and guide, only rarely assessed by statistically adequate measures!

To recapitulate: such measures comprise a comparison of the response observed (Poisson or Binomial) with that estimated by the model at each of the n levels of $\underline{x}_i^T$, by means of an aggregate statistic, the sum, $RSS = \Sigma e_i^2$, of squared residuals e_i (or e_i^* where the asterisk denotes that a chi-squared, χ_i, deviance, d_i, or Freeman-Tukey, g_i, residual is appropriate) appropriate to the distribution of the random part of the response (Binomial or Poisson) or a decrement, ΔRSS, or a function, say, AIC = RSS -(n-k), of that sum, together, of course, with some evidence that, for either the sum or decrement, the appropriate sampling distribution, e.g., Pearson's chi-squared distribution is valid - supplemented by a sum of squared predicted residuals, PRESS = $\Sigma e_{(i)}{}^2$, together with case analyses of these residuals, and of other regression diagnostics, such as Cook's distance, a standardized measure of the shift, $\hat{\underline{\beta}}_{(i)}$ - $\hat{\underline{\beta}}$, in the sample estimate of the parameter vector that is produced by the deletion of the i^{th} observation, the measures of variance inflation such as VIF, etc. The parameter estimates, $\hat{\underline{\beta}}$, and functions of the parameter estimates, such as the response, say $\underline{x}_i{}^T\hat{\underline{\beta}}$ are compared with a priori information by the Bayesian measures such as the γ-statistic described in section 7.2.4.

Moreover, "In the mature sciences the prelude to much discovery and to all novel theory is not ignorance but the recognition that something has gone wrong with existing knowledge and beliefs" (Kuhn, 1977). Comparison of the observed and expected responses by means of both case and aggregate statistics of goodness-of-fit, e.g., the chi-residuals, χ_i, and their sum of squares, $\Sigma\chi_i{}^2$, respectively, relies on methods and criteria of unrivaled authority and finesse for recognizing that "something has gone wrong" - as we will demonstrate.

However, in nearly all of the published studies listed in Table 1 there were no measures of goodness-of-fit whatever reported. The author(s) of such studies evidently assume - and require the reader to accept - that the model deployed is an adequate "fit" to the data of the report. But as Ehrenberg (1975) has remarked, it seems quite reasonable to insist that any model that is constructed largely on conjecture and speculation (for example, the LQ) be demonstrated to "fit" - on the evidence of statistically adequate measures - at least one set of empirical data (Urban's Principle of Minimum Empirical Viability).

Frequently, as in the case of the so-called LQ model of radiation toxicity by which so many investigators are now possessed, the goodness-of-fit of the model of dose-response is demonstrated by such ad hoc, surrogate - and in the event, often spurious - measures as the "straightness" of the corresponding π = 0.50 isoeffect curve, the so-called F_e-plot, the degree of "straightness" being assessed either "by eye" or by fitting a straight line to that curious transmogrification of the data, [D^{-1}, D/N], vide supra. (See Figs. 2, 10, and Fowler, 1984; Tucker and Thames, 1983). In the case of the LQ model of mutagenesis over the range of dose $0 \leq D \leq 100$ rads (Sparrow et al, 1972; NCRP 64, 1980) still another trompe d'oeil measure, the presence of not one but two straight lines with ("significantly") different slopes in the log-log plot of the dose-response (m_i, D_i) observations is cited as empirical evidence that the LQ model "fits" these data: "The apparent curvilinearity of the x-ray line can be approximated by two straight line segments, one with slope + 1.4 from about 5 to 100 rads, the other with slope +1 from 0.25 to 6 rads. A t-test indicates that the slopes differ significantly. However, a more meaningful interpretation is that the entire ascending portion of the x-ray curve can be fit as the sum of a linear and a quadratic dose term." Sparrow et al, 1972. But, the "fit" of these data to the second degree polynomial is never demonstrated by statistically adequate measures.

In some published studies of the LQ model of cell survival, still another spurious measure, the covariance matrix, Var($\hat{\underline{\beta}}$), of the sample estimate, $\hat{\underline{\beta}}$, of the parameter is interpreted as a measure of the concordance of the model and data: "For each cell line and with each model, we obtained a survival curve characterized by fitted parameters. The experimental fluctuations and the quality of the fit were expressed by both the variances and the covariance(s) linked to these parameters.... the experimental fluctuations and the quality of fitting are represented by a 95% confidence ellipse or ellipsoid ..." Fertil et al, 1980. But this procedure also simply cannot be justified: For example, if the model underfits the data, there will be "patterns" in the plot of residuals. But "pattern" is evidence of a positive correlation in the residuals, and hence that the covariance matrix, Var($\hat{\underline{\beta}}$) is deflated, As a consequence, the parameters β_j will be estimated precisely but inaccurately. The bias in $\hat{\beta}_j$ - the parameter estimates are aliased - will be large because the deterministic part of the model is misspecified.

In a few other studies the overall fit of the model is (spuriously) assessed, again "by eye", from the superpositions of sigmoid dose-response curves and the respective scattergrams of observed levels of Binomial responses. (See Tucker and Thames, 1983.)

It can be shown that these sui generis measures of "fit" provide dubious evidence of concordance. For example, the precision of the estimates of the coefficients of a model that does not "fit" the data - on statistically adequate measures - may be quite good, that is, $\beta/\sqrt{Var(\hat{\beta}_j)}$, may be large, say 3-4, for models of data for which the RSS is large enough to reject the null hypothesis that the model "fits". Or, for a given set of data, the F_e-plot may be "straight" but the LQ model does not fit the data on the evidence of the statistically adequate measures described above, i.e., comparison of the response observed with that predicted by the model, as described by the set of appropriate residuals, say χ_i, and their respective sums of squares, say $\chi^2 = \Sigma\chi_i^2$. We now proceed to examine the issue of the "fit" of the LQ model. We turn now to a closer examination of some of these examples.

7.9.1 Does the LQ model "fit" radiation toxicity data?

"The importance of producing and analyzing plots as a standard part of statistical analysis cannot be over-emphasized. Besides occasionally providing an easy to understand summary of a complex problem, they allow the simultaneous examination of the data as an aggregate while clearly displaying the behaviour of the individual cases."

A. Weissberg, 1980

"If a series of fractionated schedules have been used, so that isoeffect total doses are known for a set of dose per fraction, the ratio (α/β) can be found simply by plotting $[D^{-1}$ vs $D/N]$ yields a line whose straightness is a test of the validity of the $[\alpha D + \beta D^2/N]$ formula."

J. Fowler, 1984

"The dose-response formulae that could be used with statistical legitimacy include (1) the linear quadratic (LQ) model ... For present purposes the LQ model offers (a) valid approximations for all doses likely to be used in radiotherapy ... I shall use the $\alpha d + \beta d^2$ model because I believe it to be ... valid for doses per fraction in the radiotherapy range, i.e., up to about 10 Gy ..."

J. Fowler, 1984

"Has anything been shown to be true which even without data could not have been inferred from the assumptions?"

E. A. Murphy, 1976

There is at least one published study of the multifraction LQ model of dose-response in which the experiment is designed so that the F_e-plot is constrained to be "straight", no matter whether the model "fits" the data - on statistically adequate measures and criteria (e.g., the residuals and their sum of squares) - or not; the experiment becomes a mere tautology or "self-fulfilling prophecy". As this study is discussed at length in Annex II, part 3, we present only a brief exposition here based on the data of Figs. 9 and 10a which are taken from the published study in question (Tucker and Thames, 1983).

The data of Fig. 9 have been replotted in the four different scattergrams of Figs. 14a, 14b, and 15a, 15b. Figure 14a presents a scattergram in the (D, D/N)-plane of the 19 treatment regimens, $\underline{x}_i^T$, of Fig. 9. The filled symbols denote those 7 regimens for which $0 < r_i < n_i$, where r_i is the number of responders out of n_i at risk at $\underline{x}_i^2$. The open symbols denote those regimens at which the response is extreme: $r_i = 0$ (seven) or $r_i = n_i$ (five). Thus, an excess, 63%, of the data of Fig. 9 resides in these extreme responses. Note that in 9 of the 19 regimens, the dose per fraction d = D/N exceeds 10 Gy, the level of d that is stipulated as the upper limit of the range of validity of the LQ model. (Fowler 1984, 1989) Moreover, in 19 of the 19 regimens d lies outside the range of, "doses per fraction in the radiotherapy range" that is described by the cross hatched region of Fig. 14a.

The distribution of observations in Fig. 14a is concave up. Hence, it should come as no surprise to find that it is "straightened" by the transformation $D \longrightarrow D^{-1}$. Figure 14b presents a

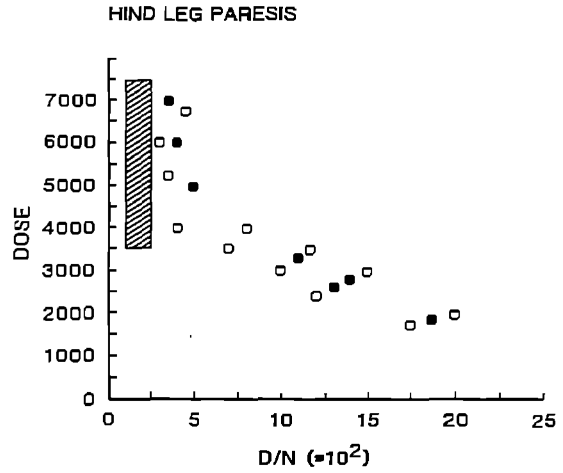


Fig. 14a. The plot is a scattergram in the (D, D/N)-plane of the observations of the designed experiment described in Fig. 9. The filled symbols identify those levels of treatment (D_i, D_i/N_i) for which $0 < r_i < n_i$. The open symbols identify those levels for which either $r_i = 0$ or $r_i = n_i$. The cross-hatched area identifies the region of the (D, D/N)-plane corresponding to clinical treatment regimens ("the radiotherapy range"). Note i) that all of the 19 regimens lie outside the clinical region - an example of looking where the light is better rather than where the key might lie; ii) that 9 of the 19 regimens lie beyond the stipulated range of d = D/N = 10 Gy for which the LQ model is valid; iii) that an excess of the data lies in the extreme reponses, $r_i = 0$ or $r_i = n_i$; iv) that the distribution of the observations is concave-up. (N.B. "I shall use the $\alpha d + \beta d^2$ model because I believe it to be both valid for doses per fraction in the radiotherapy range; i.e., up to about 10 Gy ..." (Fowler, 1984). (Emphasis added)

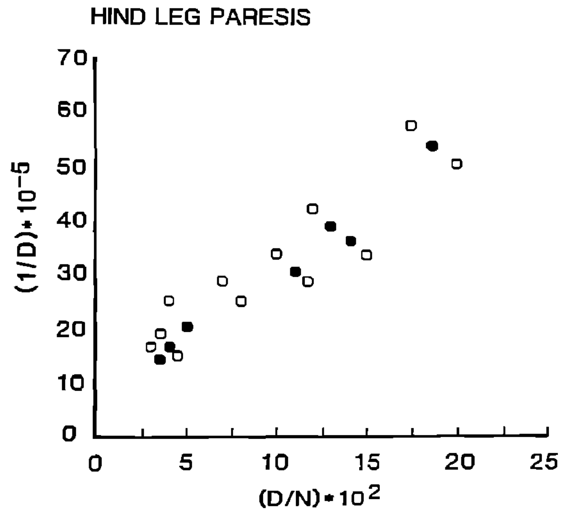


Fig. 14b. The plot is a scattergram in the (D^{-1}, D/N)-plane of the designed experiment described in Fig. 9. Note that since the plot of (D, D/N) is concave-up in Fig. 9 it is not surprising that the plot of (D^{-1}, D/N) is a straight band. The equation of the line that best graduates the 19 observations is nearly identical with the equation of the F_e-plot of Fig. 10, suggesting that the F_e-plot will be "straight" whether the LQ model "fits" - on normative criteria (comparison of the observed and predicted responses) - those observations on dose-response or not. Therefore, the degree of "straightness" of the F_e-plot would seem to be a very poor criterion of the validity of the multifraction LQ model for the data at issue. (N.B.: "When you have something that one of the gods wants you to discover you certainly do not have to do a statistical test. No one has to prove the linearity of the data. There it is. By god, it is linear. With such a discovery in hand, statistical tests become irrelevant. And irreverent" (R.C. Bolles, 1988).)

Although it is surely no surprise, on the evidence of Fig. 14, that the = 0.50 isoeffect plot of Figure 10 is indeed "straight", it is altogether quite astonishing that the proposition that the goodness-of-fit ("validity") of the multifraction LQ dose-response model, $E = \alpha D + \beta D^2/N$, could be adequately assessed by any feature of the cognate isoeffect model, $D^{-1}(\pi) = \alpha + D(\pi\ )/N$, should be so widely accepted. But it is of course the case, that it is also widely accepted that the ratio, α/β, which can only be unique to within a multiplicative constant, uniquely characterizes the radiation response of a tissue so unambiguously that it can be deployed as a discriminant criterion, perhaps such acceptance is less remarkable than it first appears.

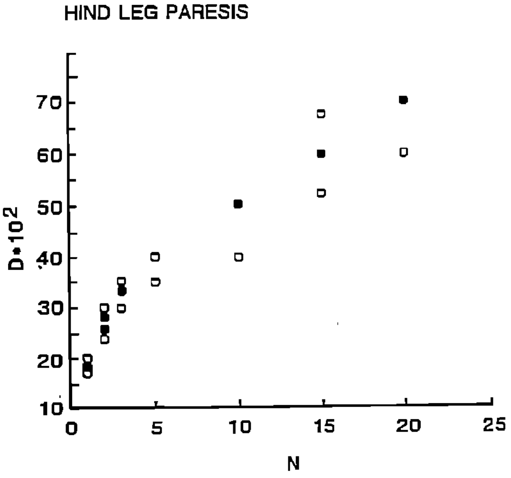


Fig. 15a. The plot is a scattergram in the (D,N)-plane of the observations of the designed experiment described in Fig. 9. Note that the distribution of the observations is concave-down.

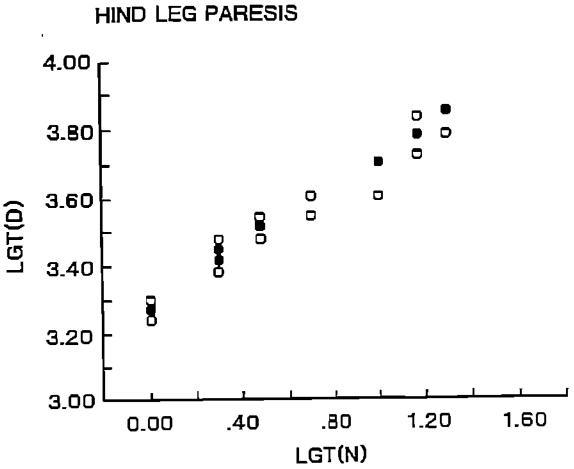


Fig. 15b. The plot is a scattergram in the $(\log_{10}D, \log_{10}N)$-plane of the observations of the designed experiment described in Fig. 9. Note that the distribution is, as in the case of Fig. 14b, a straight band; since the distribution of the observations in the (D,N)-plane is concave-down this comes as no surprise. Obviously, any <u>isoeffect</u> curve, log D(x) vs log N must also be a sraight line suggesting that a model in which the treatment variables are x1 (= logD) and x_2 (= logN) is equally as valid as the LQ model.

scattergram of the transformed treatment regimens: (D, D/N) ⟶ (D^{-1}, D/N). The set of 19 regimens is graduated very well by the equation (least squares estimates), $D^{-1} = 0.095 + 0.217(D/N)$; $R^2 = 0.862$. This equation is quite similar to the equation (least squares estimates) that graduates the F_e-plot of Fig. 10: $D^{-1}(\pi) = (\alpha/E) + (\beta/E)D(\pi)/N = 0.09 + 0.218D(\pi)/N$; $R^2 = 0.982$. Here $\pi = 0.50$ is the proportion of responders in the 50% isoeffect subset. Therefore, the "straightness" of the F_e-plot that "validates" the LQ model appears to be built-in to this experiment; the design of the experiment has effectively "immunized" the data against any "lack of fit" by the LQ model. Whatever may be the agreement - or lack thereof (vide infra) - between the level of response observed, r_i responders in n_i exposed at $\underline{x}_i^T$, $1 \leq i \leq n = 19$, and that predicted by the LQ model, $n_i\hat{\pi}_i$, the cognate isoeffect curve is constrained to be a straight line by the curious design of this experiment. For this experiment the criterion of "straightness" of the F_e-plot is merely a tautology: it is entailed in the experimental design.

Figure 15a presents a scattergram of the treatment regimens, $\underline{x}_i^T$, in the (D,N)-plane. The distribution of observations is concave down. Hence, it should surprise no one that it can be "straightened" by the transformation D ⟶ logD, N ⟶ logN. Figure 15b presents a scattergram of the logarithmic transform of the 19 treatment regimens x_1 = logD, x_2 = logN plane. The set is graduated very well by the equation $x_1 = 3.297 + 0.397x_2$; $R^2 = 0.945$. Clearly, this empirical transformation could describe the isoeffect subset, $(x_1(\pi), x_2)$ for a model of dose-response that has, on the (received) criterion of "straightness of the 0.50 isoeffect curve", at least as much "validity" as has the LQ model.

For the experiment described in Fig. 9 the response is quantal and hence has a Binomial distribution. A link function for the appropriate generalized linear model of these data is the probit (or logit). It also appears that the response depends upon the number of fractions N and perhaps the time T. The rival dose-response models cognate to the isoeffect models of Figs. 14b and 15b are:
a) $z = \beta_0 + \beta_2 D + \beta_2 D^2/N$ (multifraction LQ model). b) $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (Power-law).
Here the probit transform is $z = \Phi^{-1}(\pi)$, $0 \leq \pi \leq 1$, $\Phi(.)$ is the Normal distribution function and $x_1 = \log D$, $x_2 = \log N$. These are designated as models M2 and M1b, respectively (in Annex II). The respective sums of chi-squared residuals are RSS $= \Sigma e_i^{*2} = \chi_c^2 = \Sigma\chi_i^2 = \underline{41.036}$ (M2) and $\chi_c^2 = \Sigma\chi_i^2 = \underline{42.686}$ (M1b). The number of degrees of freedom for each is (n-k) = 16. The data reject both models:

$$P(\chi^2 > \chi_c^2|16) = 3*10^{-4} \text{ (M2) and } P(\chi^2 > \chi_c^2|16) = 3.8*10^{-4} \text{ (M1b).}$$

However, respective residual sums of squares, RSS $= \Sigma e_i^{*2}$, for the rival models, may, because many expected frequencies are too small (a consequence of the small numbers at risk, n_i) have a sampling distribution that is not well-approximated by any well-known distribution, such as Pearson chi-squared; therefore, supplementary measures of concordance must be considered. Both the deviance and Pearson chi-squared measures - Σd_i^2 and $\Sigma\chi_i^2$, respectively - may overstate the degree of departure from the fitted model when many cells contain small counts (Fienberg, 1980) and when they are interpreted as chi-squared statistics; a correction factor for deviance is available (Williams, 1976). Hosmer and Lemeshow (1989) describe a grouping procedure based on percentiles of the estimated probabilities, $\hat{\pi}_i$, ("deciles of risk") that can be deployed when the numbers at risk are as small as $n_i = 1$, $1 \leq i \leq n$.

Of course, even if the chi-squared approximation is valid, the RSS provides only an aggregate statistic for a summary test of goodness-of-fit and, as Robins and Greenland (1986) - and others - have remarked, the "... problem with goodness-of-fit tests is that they are insensitive to certain inconsistencies between models and data, and so they can indicate a good fit for models that are grossly inconsistent with the data. Thus, if one wishes to use a model consistent with the data, one ought to check for model adequacy by examining residuals, screening for outliers, and employing other similar diagnostic techniques. But even among the class of models consistent with the data, the intermodel variation in the estimated exposure effect may exceed the model specific standard errors." And, "... residuals are mainly used to look for what might be going on beyond what is already in the model ... The important thing is to look at the residuals ..." J. Tukey (1968).

Thus, one must always examine the several scattergrams, e.g., e_i^* vs $\hat{\mu}$, $1 \leq i \leq n$, of the individual residuals, in the present case, the chi-residuals, $e_i^* = \chi_i$ vs $\hat{\pi}_i$, of a given model of a given set of data. In particular, we must note whether any of the distributions exhibits a "pattern" and also whether there are large e_i^* that identify responses that are not well explained by the model, i.e., whether there are any "outliers", as well as whether there are any observations that <u>dominate</u> either the "fit", or the estimates, $\underline{\hat{\beta}}$ and $Var(\underline{\hat{\beta}})$.

The respective plots of e_i^* vs $\hat{\pi}_i(= \bar{P}_i)$ for models M1b and M2 are presented in Figs. 16a and 16b. (<u>N.B.</u> For data with binomial errors, note that $\hat{\mu}$, the conditional response estimated from the model, is interpreted as $\hat{\pi}$ rather than $n\hat{\pi}$ (McCullagh and Nelder, 1989).) Each plot displays the "double-bow" characteristic of a Bernoulli variate for which the variance is a <u>maximum</u> at $\pi = 0.50$. (Montgomery and Peck, 1982). The plot for M1b also discloses the presence of two "outliers": $3.0 < |e_i^*|$. These two, $e_6^* = -3.58$ and $e_{14}^* = 3.03$, describe a "pattern" - a straight line. The plot for M2 discloses the presence of two "outliers": $3.0 < |e_i^*$, $i = 6, 14$. These two are more extreme $e_6^* = -3.74$ and $e_{14}^* = 3.79$ and describe the same "pattern" - a straight line. (Note that the observations that define the "linear" pattern of the residuals, #6 and #14, are <u>interior</u> to the sample and are not otherwise extreme or atypical.) Thus models M1b and M2 are obviously misspecified: They have not <u>captured</u> all of the information in those data - some has "leaked" into the residuals (Box, Hunter and Hunter, 1978).

Figure 16c presents a plot of the residuals, e_i^*, for the model M2 vs e_i^* for M1b. The extreme points represent the respective residuals for cases #6 and #14. The correlation between the respective sets of residuals for the two models is quite high: $r = 0.944$. Thus, it appears that these two models, each a function <u>only</u> of dose, D, and number fractions, N, <u>fail</u> by approximately similar amounts, RSS = 41.036(M2) vs RSS = 42.686(M1b), <u>and in a similar manner</u>, i.e., that each model is misspecified in a similar way. For example, neither includes the time factor.

Figure 16d is a two-sample Q-Q plot in which the respective quantiles of the distribution of the chi-residuals for model M2 (LQ) are plotted against the cognate quantiles of the distribution of chi-residuals for model M1b. The plot is nearly straight suggesting that the two samples could be from the same distribution. This plot reinforces the evidence of Fig. 16c above, that the models M1b and M2 (LQ) fail to "fit" the data of Fig. 9 by similar amounts and in similar ways. But, note that although neither of the models M1b nor M2(LQ) "fit" the data of Fig. 9, the respective parameter estimates are highly significant: $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)} > 7.0$, $0 \leq j \leq 2$. This is a characteristic "strength" of models that <u>underfit</u> the sample data and suggests that the recommendation of Fertil et al (1980) that a high precision of the parameter estimates is evidence of "the quality of the fit" (of the model) is <u>not</u> "robust".

We have seen in Fig. 2a that the sample estimate of α/β for the LQ model of the dose response of mouse jejunum crypt cell (Fowler, 1984) obtained from the F_e-plot is <u>not</u> robust either: deletion of the observation at D/N = 15 Gy <u>inflates</u> the estimate of α/β by a factor of two. We have shown above that the assessment of the goodness-of-fit of the LQ model of dose-response of rat spinal cord (hind-leg paresis) to the data of Fig. 9 on the criterion of the "straightness" of the Fe-plot of Fig. 10a is misleading. It would seem that the deployment of the typical F_e-plot may well yield both <u>unfounded inferences</u> (on goodness-of-fit) and <u>unstable estimates</u> (of α/β).

The presence of a linear "pattern" in each of the residuals plots, e_i^* vs $\hat{\pi}_i$, of models M1b and M2 suggests that the respective "fits" could be improved by including, in a revised model, a first-order term in one of the treatment variables. An obvious choice is the duration of the treatment schedule, T - or a function thereof, say $x_3 = \log T$, as suggested by Fig. 17. Figure 17 is a so-called "added-variable plot" that provides a graphical measure of whether an omitted covariate, here T, could be usefully included in the deterministic part, η, of the model.

It is useful at this point to recur to the definition of a "statistically adequate" model (as distinct from a "statistically legitimate" one). A model, $F(X, \underline{\beta})$, of a set of n observations, $[\underline{y}, X]$, is statistically adequate if it is <u>plausible</u> that a parameter vector $\underline{\beta}$ exists such that $(\underline{y} - F(X, \underline{\beta}))^T = \underline{e}^T = (e_1, e_2, ..., e_n)$ is a noise sequence unrelated to any known variable. (Box, Hunter and Hunter, 1978). Here y_i, $1 \leq i \leq n$, is the observed value of the <u>response variable</u> at the i^{th} row,

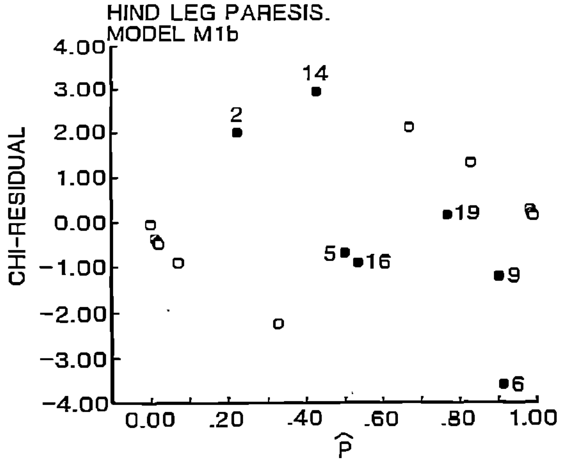


Fig. 16a. Plot of the chi-residuals $\chi_i = (r_i - \hat{\pi}_i n_i)/\sqrt{n_i \hat{\pi}_i (1-\hat{\pi}_i)}$, vs $\hat{P}(= \hat{\pi}_i)$ for the probit model M1b with linear predictor, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ of the data of the experiment described in Fig. 9. The filled symbols identify those regimens for which $0 < r_i < n_i$, the open symbols those for which $r_i = 0$ or $r_i = n_i$. Adjacent to each of the seven regimens for which $0 < r_i < n_i$, is the <u>index number</u> of the observation. There is a double-bow pattern characteristic of a Bernoulli variable for which the variance is greatest at $\pi = 0.50$. There are also two outlying observations for which $\chi_i \cong 3.0$. The aggregate goodness-of-fit measure is $\chi_c^2 = \text{RSS} = \Sigma\chi_i^2 = 42.69$ which is distributed as χ^2 on $(n-k) = 16$ degrees of freedom (df) and $P(\chi^2 > \chi_c^2|16) = \underline{4{*}10^{-4}}$. This suggests that the model M1b should be <u>rejected</u> on the normative (statistical) criteria of goodness--of-fit.

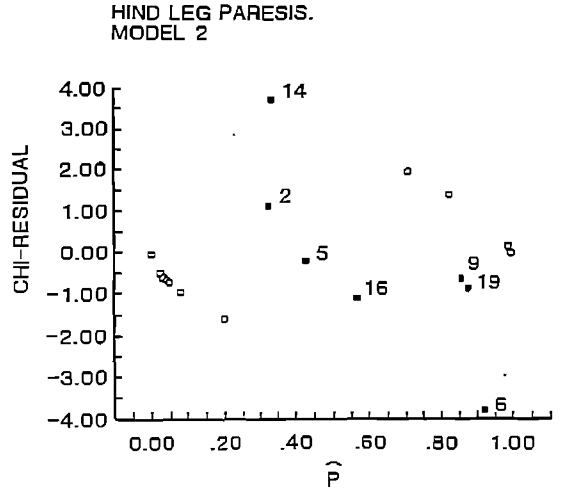


Fig. 16b. Plot of the chi-residuals $\chi_i = (r_i - \hat{\pi}_i n_i)/\sqrt{n_i \hat{\pi}_i (1- \ i)}$ vs $\hat{P}_i(= \hat{\pi}_i)$ for the probit model M2 with linear predictor, $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N$ of the data of the experiment described in Fig. 9. The symbols and numbers are as in Fig. 16a. The linear predictor for model M2 is that <u>cognate</u> to the <u>multifraction</u> ($N \geq 1$) LQ model: Effect = $\alpha D + \beta D^2/N$.

There is the double-bow pattern characteristic of a Bernoulli variable in which the variance is greatest at $\eta = 0.50$. There are two <u>outlying</u> observations. The <u>aggregate</u> goodness-of-fit measure is $\chi_c^2 = \text{RSS} = \Sigma\chi_i^2 = \underline{41.036}$ which is distributed as χ^2 on $(n-k) = 16$ degrees of freedom (df) and $P(\chi^2 > \chi_c^2|16) = 3{*}10^{-4}$. This suggests that the model M2 should be <u>rejected</u> on the normative criteria of goodness-of-fit.

The residuals plots for models M1b and M2 suggest that neither of these models, in which the response is a function only of D and N, have captured all of the information on dose-response that was included in the data of the Tucker and Thames (1983) experiment described in Fig. 9; some has obviously "leaked" into the residuals. Note that the <u>sample size</u>, n=19 > k + 15 and hence is large enough to support "... a meaningful residual analysis" (Snee, 1977).

$\underline{x}_i^T$, of the model matrix X. Note that $\underline{e}^{T*}\underline{e} = \Sigma e_i^2$ = RSS, the sum of squared residuals, the usual statistic whose sampling distribution determines the concordance of model and data. Thus, knowledge of both the "size" and "shape" (or, better, lack thereof, vide supra) of the scattergrams (e.g., e_i vs x_i) of $\underline{e}$ are required to determine the statistical adequacy of the model, F(X, $\underline{\beta}$).

"Noise", or better, "white noise", describes a distribution of e that is featureless, without any regularities or pattern. An absence of "pattern" in the residuals implies that all of the information on the conditional response has been recovered by the model. A useful picture is that a statistically adequate model "maps" the data into a white noise vector: $\underline{\beta}$:[$\underline{y}$, X] $\underline{e}$. Note that the statistical adequacy of the LQ model of a given multifraction data set can seldom be determined since the response variable (cell survival, S) described by that model is rarely identified, and often cannot be observed (and its relation to any proxy variable that could be observed is never specified).

Augmenting the model M1b by including the variable x_3 = logT reduces RSS by 50% and gives the model M1a:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.\ \text{RSS} = \chi_c^2 = \Sigma\chi_i^2 = 21.93.\ P(\chi^2 > \chi_c^2|15) = 0.109.$$

The data do not reject this model on the aggregate measure, RSS. Note also that the decrement, ΔRSS = 42.69 - 21.93 = 20.76, between the nested models M1b and M1a, which is distributed as χ^2 on 1 df, is obviously "highly significant". Note that the sampling distribution of the decrement, ΔRSS, is well-approximated by the chi-squared distribution on $k_{1a} - k_{1b}$ = 4-3 = 1 degree of freedom, $\chi^2(1)$, even though the sampling distributions for the respective values of RSS for models M1a and M1b may not be well-approximated by the cognate chi-squared distributions. Moreover, we find (see Annex II) that each of the four parameter estimates exceeds its standard error significantly: $|\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}| \geq 4.0$, $0 \leq j \leq 3$. However, we also find that $\beta_3 < 0$. This is either a "novelty of fact", or, what is more likely the case, an artifact due to the large correlation of N and T in these data as is evident in Fig. 3. A more complete exposition is given in Annex II, part 2. Of course, it is commonly believed that the predictive performance of a regression model cannot be increased by adding a predictor variable which is highly correlated with one or more of the predictor variables already included in the model. However, Cramer (1974) and Kotz and Johnson (1982) show that this is not necessarily the case and, in fact, has no basis in statistical theory. Both Cramer (1974) and Kotz and Johnson (1982) give examples in which two predictor variables are very highly correlated and yet each contributes appreciably to the predictive performance of the model. Model M1a apparently provides yet another example.

We have repeatedly (and quite correctly) criticized as an invalid inferential procedure, the interpretation of the "straightness" of the F_e-plot as a test of the goodness-of-fit of the multifraction LQ model. However, it must be noted that there are two special cases in which the capacity of the human eye to readily discriminate, in a bivariate plot, between a line that is straight and one that is not, can be usefully exploited as an important adjunctive graphical test of a hypothesis. The basic principle in these circumstances is that selected statistics of the data are plotted on a grid for which the scales of the two coordinate axes are so chosen that if the data are not inconsistent with the hypothesis, then the scattergram of statistics can be well-described by a straight line; that is, they will lie close to, and randomly scattered about, a straight line. Any systematic deviation from a straight line indicates that the data reject the hypothesis - and often gives an indication of the reason for the rejection (Gilchrist, 1984). These graphical procedures may be deployed adjunctively to test hypotheses on both the form of the deterministic part (μ_i) and the form of the distribution of the random part (ϵ_i) of the response, $y_i = \mu_i + \epsilon_i$, $1 \leq i \leq n$. In fact, it is a common practice to do so. (Montgomery and Peck, 1982) One acceptable graphical procedure is to plot the observed, y_i, vs expected, $\hat{\mu}_i$ (or $\hat{y}_i$), levels of response to test hypotheses on the form of the deterministic part. Another is to plot the order statistics of the standardized residuals e_i^* (where $e_i^* = (y_i - \hat{\mu}_i)/\sigma(1-h_i)$ for a Normal theory model) against the quantiles of the proposed distribution (say $z_i\{i/(n+1)\}$ where $z = \Phi^{-1}(\pi)$, $0 \leq \pi \leq 1$) to test hypotheses on the form of the distribution of the random part; i.e., a probability plot. (We have previously illustrated the use of probability plots in Figs. 5a and 6a.) However, each graphical test should be supplemented by an examination of the appropriate correlation coefficient (for the Normal

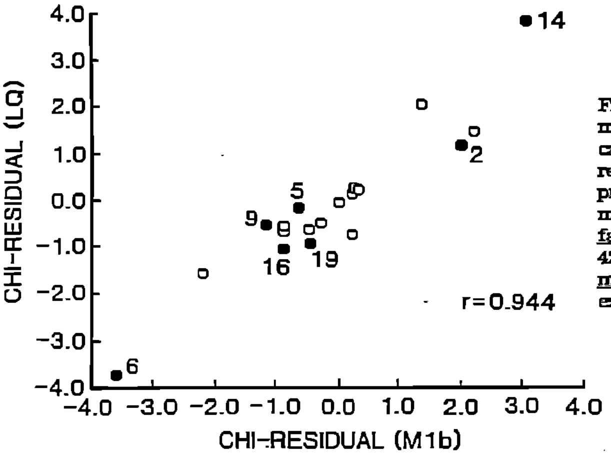


Fig. 16c. Plot of the chi-residuals, χ_i, for the model M1b vs χ_i for the model M2. The extreme points represent the respective residuals for cases #6 and #14. The correlation between the respective sets of residuals for the two models is quite high: r = 0.944, where r is the product moment correlation coefficient. Thus, it appears that these two models, each a function only of dose, D, and number of fractions, N, fail by approximately similar amounts, RSS = 41.036 (M2) vs RSS = 42.686 (M1b), and in a similar manner, i.e., that each model is misspecified in a similar way (Neither includes a time factor, for example).

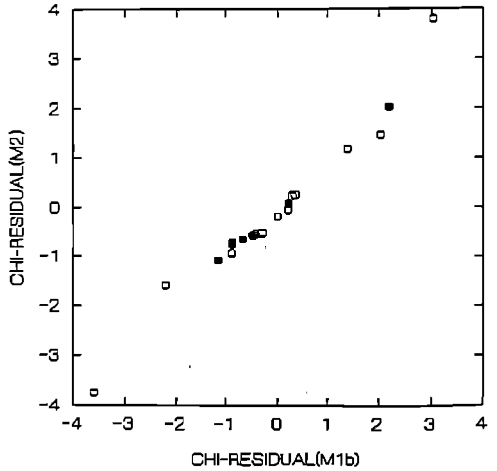


Fig. 16d. The figure is a two-sample Q-Q plot in which the quantiles of the distribution of the chi-residuals for model M2(LQ) are plotted against the cognate quantiles of the distribution of the chi-residuals for model M1b. The plot is nearly straight suggesting that the two samples could be from the same distribution. This reinforces the evidence of Fig. 16c above, that the M1b and M2(LQ) models of the data of Fig. 9 fail by similar amounts and in similar ways.

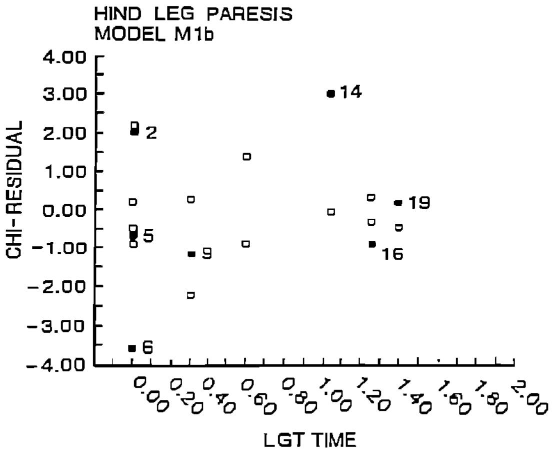


Fig. 17. The presence of a linear "pattern" in each of the plots of the chi-residuals, χ_i vs $\hat{P}_i(=\hat{\pi}_i)$, for models M1b and M2 suggested both models underfit the data and thus, that the respective "fits" of each could be improved by including, in a revised model, a first-order term in one of the predictor variables. An obvious choice is T - or a function thereof, say x_3 = logT, as is suggested by this plot.

If the linear predictor of model M1b is augmented by x_3 to give probit model M1a, with linear predictor, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, the decrement in χ^2 is highly significant: $\Delta\chi^2 = 42.69 - 21.93 = 20.2076 > 3.84$. (The 0.95 quantile of the distribution of χ^2 of 1 df is 3.84.) Moreover, the data do not reject the model M1a since $\chi_c^2 = 21.93$, $P(\chi^2 > \chi_c^2|15) = 0.109$. Note that although the validity of the chi-squared approximation may be questionable for the measure of overall fit, χ_c^2, owing to many small expectations, that is not the case for the decrement, $\Delta\chi^2$.

On the other hand if the linear predictor of model M2, the multifraction LQ model, is augmented by β_3T to give model M3, $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3 T$, the decrement in chi-squared is insignificant: $\Delta\chi^2 = 41.036 - 41.036 = 0$. If the linear predictor is augmented by x_3 = logT to give the model M4, $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3$ logT, the decrement in χ^2 is significant: $\Delta\chi^2 = 41.036 - 27.688 = 13.35$. However, the data reject the model M4: $\chi_c^2 = 27.688$, $P(\chi^2 > \chi_c^2|15) = 0.024$. (This points up the fact that the better-fitting of two rival models still may not fit the data very well; hence, there is a requirement to examine the overall fit of every model.)

distribution this is the Filliben probability plot correlation coefficient. (Filliben, 1975).)

The plot of observed, P_i, vs expected, $\hat{P}_i$, responses is shown in Fig. 18a for the model M1a of the Binomial data of Fig. 9. The dashed line is the line of perfect fit. The fit seems adequate, save for the observation 14, which appears to be an "outlier".

A probability plot of the residuals for M1a is presented in Fig. 18b. The evident linearity of this plot together with the evidence of the Filliben (1975) probability plot correlation coefficient, $r_F = 0.961$, suggests that the inferences on goodness-of-fit, based on the assumption of a Pearson chi-squared distribution for the RSS of model M1a, should be fairly good, despite the rather large number of observations at which the expected frequency is very small indeed; it will be recalled that in 12 of the 19 observations, either $r_i = 0$ or $r_i = n_i$. Moreover, "In linear regression a probability plot with non-linear shape is generally taken to imply that the error distribution is not modelled correctly. One does not hope to learn anything about the adequacy of the $X\hat{\underline{\beta}}$ term from a probability plot. In logistic regression the situation is different. Because of the discrete nature of the response with expectation p and variance p(1-p), adequate modelling of the entire error distribution is inherently bound up with adequate modelling of the expectation term $X\hat{\underline{\beta}}$. Thus, differences that appear in an empirical probability plot from logistic regression might be related to an inadequacy in $X\hat{\underline{\beta}}$ ..." (Landwehr, Pregibon, and Shoemaker, 1984). A similar remark obtains, of course, for probit regression models. Thus, in the interpretation of the plots of Figs. 18a and 18b, the degree of "straightness" of the plots is a valid criterion for adequacy of "the match" of theory and data: Figure 18a, a fit-observation plot, provides a measure of the adequacy of the selected model of the deterministic part, $\underline{\mu}$, of the response; Figure 18b, a Normal probability plot, provides a measure of the adequacy of the selected model of the form of the distribution of the stochastic part, $\underline{\varepsilon}$, of the response (and also, as remarked above, of the model of the form of $\underline{\mu}$). See also Figs. 24a and 24b. However, as shown previously, that in the interpretation of the so-called F_e-plot of Fig. 10a, the degree of "straightness" is a spurious criterion of the match.

As we have found for the models M1b and M2 of the data of Fig. 9, it is also instructive to plot the chi-residuals of the model M1a against the expected levels of response. The chi-residuals may be plotted either against the fitted values $\hat{\pi}_i$ "... or against the fitted values transformed to the constant information scale of the error distribution" (McCullagh and Nelder, 1989). For Binomial errors the latter scale is $2\sin^{-1}\sqrt{\pi}$. Such a plot is presented in Fig. 18c. There is the characteristic "double-bow" seen in Figs. 16a and 16b and a single (possible) outlier, $e_{14}{}^{*} = 2.64$, superimposed. The absence of any pattern suggests that this model has captured all of the information on dose-response that is contained in the experiment of Fig. 9. Note that it is a characteristic feature of all residual-vs-fitted-value plots that observations with the same level of response lie on parallel straight lines with negative slope (Searle, 1988).

The sample estimates of the parameter vector $\underline{\beta}$ for model M1a are quite stable: $\bar{\beta}_j/\sqrt{Var(\bar{\beta}_j)} \geq 7.0$, $0 \leq j \leq 2$; $\bar{\beta}_3/\sqrt{Var(\hat{\beta}_3)} > 4$. These are very near the levels required for a "useful model": $\hat{\beta}_j/\sqrt{Var(\bar{\beta}_j)} > 8.0$, $0 \leq j \leq p$, as we have discussed earlier.

Although it would appear from the evidence of both the aggregate and case statistics of fit that the model M1a is a statistically adequate model of the data of Fig. 9, further examination of other diagnostics that are presented in Figs. 18d, 18e, and 18f discloses that this is not the case: These figures show that the sample estimates of the parameters are exceedingly labile owing to the weaknesses in the experimental design, chiefly the high degree of correlation between N and T that is shown in Fig. 3 and the small numbers, n_i, at risk at each level of treatment.

In a few other studies that we reviewed, decrements in aggregate statistics such as deviance or chi-squared are used to discriminate between rival models within a hierarchy, or "nest" of models. Of course, basing an inference of "fit" on decrements of, say, deviance between rival nested models exposes the investigator to a logical fallacy, since the model that "fits" the better of the two compared may still not fit the data very well - or, indeed, at all. See for example, Herbert 1985a.

Models M1a and M1b are nested rivals. Models M1a and M2 are non-nested rivals. We have found that the published radiobiological literature offers no instance in which non-nested, rival dose-response models of a sample are correctly compared on the criterion of concordance. However,

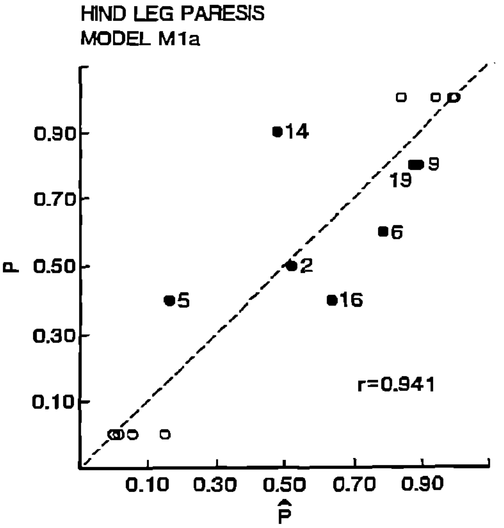


Fig. 18a. Plot of the observed response, $P_i = r_i/n_i$, vs the predicted response, $\hat{P}_i = \Phi(\hat{z}_i)$, for the probit model M1a of the data of Fig. 9. The index numbers of those 7 (of 19) treatment regimens at which the response was not extreme - $0 < r_i < n_i$ - are shown. The points appear to be scattered at random about the (dashed) line of perfect "fit": $P_i = \hat{P}_i$, $1 \leq i \leq n$ (= 19). The correlation coefficient, r = 0.941 of P_i and $\hat{P}_i$ is satisfactorily large, but is obviously dominated by the extreme responses at $P_i = 0$ and $P_i = 1.0$.

The point most remote from the line $P_i = \hat{P}_i$ is at index number 14 for which $P_i = 0.48$, $\hat{P}_i = 0.90$ ($D_{14} = 5000$ cgy, $N_{14} = 10$, $T_{14} = 11$ days). However, the variance of a Binomial variate is greatest at $\pi = 0.50$; and therefore, evaluated in this context, the large discrepancy does not suggest a serious lack of fit. However, the plot does emphasize the recommendation that aggregate statistics, such as the correlation coefficient (or its square, $r^2 = 0.885$) must, also, not be overread.

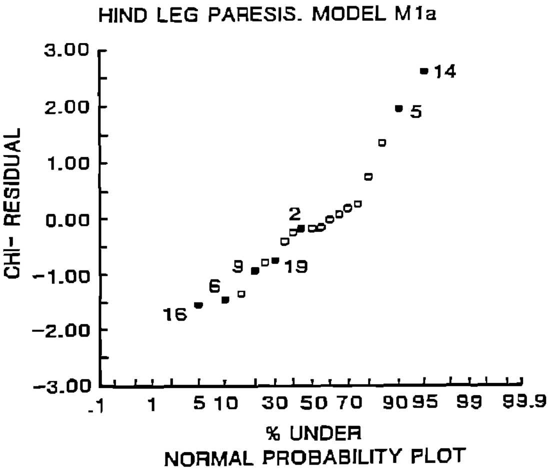


Fig. 18b. Normal probability plot of the chi-squared residuals, $\chi_i = (r_i - n_i\hat{\pi}_i)/\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$, where $\hat{\pi}_i(=\hat{P}_i) = \Phi(\hat{z}_i)$ for the probit model M1a of the data in Fig. 9. The Filliben probability plot correlation coefficient is $r_F = 0.961$. On the evidence of the "straightness" of the plot itself, together with the value of r_F, it may be concluded that the data do not reject the model. The data also do not reject the hypothesis of Normality on the Kolmogorov-Smirnoff (Lilliefors) test at p = 0.05. Note that this is a valid graphical test of goodness-of-fit, in which the (degree of) "straightness" of the plot provides valid empirical evidence for the fit of the model.

"In linear regression a probability plot with non-linear shape is generally taken to imply that the error distribution is not modelled correctly. One does not hope to learn anything about the adequacy of the $X\hat{\beta}$ term from a probability plot. In logistic regression the situation is different. Because of the discrete nature of the response with expectation p and variance p(1-p), adequate modelling of the entire error distribution is inherently bound up with adequate modelling of the expectation term $X\hat{\beta}$. Thus, differences that appear in an empirical probability plot from logistic regression might be related to an inadequacy in $X\hat{\beta}$ ..." Landwehr, Pregibon, and Shoemaker, 1984.

The "straightness" of the so-called F_e-plot is regularly offered as evidence for the goodness-of-fit of the LQ model. However, it is an invalid graphical test.

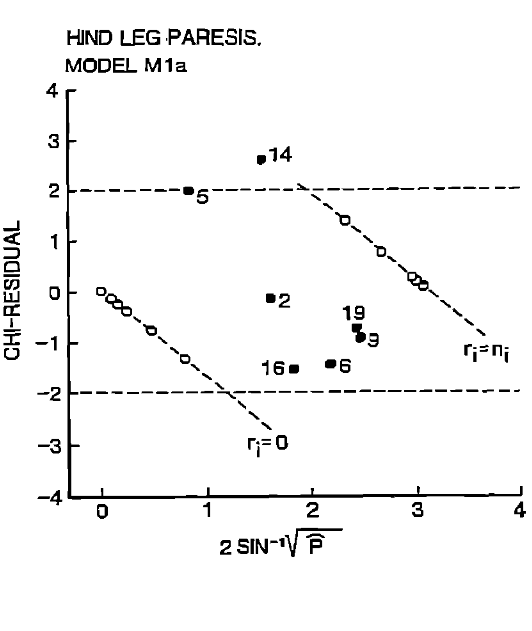


Fig. 18c. Plot of the chi-residuals $\chi_i = (r_i - \hat{\pi}_i n_i)/\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$ vs $2\sin^{-1}\sqrt{\hat{\pi}}$ for the probit model M1a with the linear predictor, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ of the experiment described in Fig. 9. The symbols and numbers are as in Fig. 16. The linear predictor for model M1a is that of an "NSD-type" model.

The basic double-bow pattern that is characteristic of a Bernoulli variable, as in Figs. 16a and 16b has been replaced by the two parallel lines with negative slope at $r_i = 0$ and $r_i = n_i$. The transformation $\hat{\pi}_i \longrightarrow 2\sin^{-1}\sqrt{\hat{\pi}_i}$ corresponds to the constant-information scale for the binomial error distribution (McCullagh and Nelder, 1989). On this scale the contours of fixed y(=r) are parallel straight lines if the form of the distribution of errors is binomial - as it appears to be in this case. There are no outlying observations suggesting that this model which includes a time-factor-, $x_3 = \log T$, has retrieved all of the information on dose-response that was contained in the data of the experiment described in Fig. 9. This evidence of the residuals plot - one of the case statistics of this model of this data - is consistent with that of the aggregate statistic $\chi_c^2 = RSS = \Sigma\chi_i^2 = 21.93$; $P(\chi^2 > \chi_c^2|15) = 0.109$. (N.B.: "It is surprising that one of the main factors hindering the acceptance of the LQ approach to working out isoeffect doses has been the alleged absence of a 'time factor' in the LQ formula. No time factor should be present for late effects" J. Fowler, 1989.) Note, however, that although the model M1a is not inconsistent with the sample data - on the evidence of case statistics, χ_1, as well as the aggregate statistic $RSS = \Sigma\chi_i^2 = 21.93$, it will not predict well in new data; RSS is a biased estimate of prediction error. A reduced bias estimate of prediction error is $PRESS = \Sigma\chi_{(i)}^2 = \Sigma\chi_i^2/(1-h_i) = 31.61$. The PRESS statistic is a jackknife validation measure. Minimum PRESS is a model discrimination criterion. The PRESS statistic can be deployed in model discrimination as well as validation. It can be used to discriminate between non-nested models (M1a vs M2) as well as nested models (M1a vs M1b), whereas the decrements $\Delta RSS = \Delta\chi^2$ or $\Delta RSS = \Delta Deviance$ cannot. In using PRESS as a discrimination criterion, the rule is to select the model with minimum PRESS: It can be shown that, on the minimum PRESS criterion, model M1a is preferred over M2.

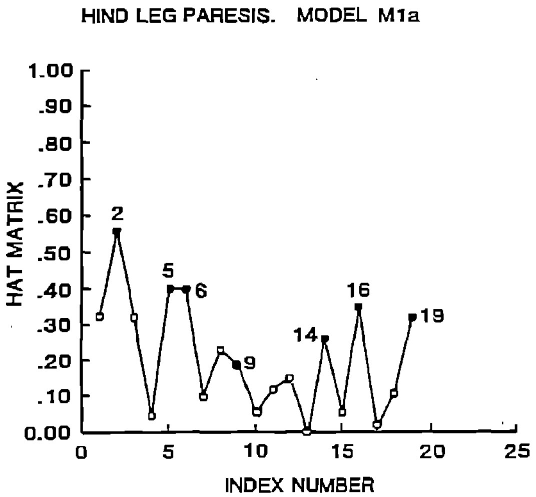


Fig. 18d. Index plot of the diagonal elements, h_i, of the (nxn) hat matrix. $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ for the generalized linear model, M1a, of the data of the experiment of Fig. 9. W is the (nxn) diagonal weight matrix for the probit transform of the Binomial response (See Cook and Weisberg, 1983). The hat matrix diagonals, h_i, together with the weighted residuals, $e_i^* = \chi_i$, of Fig. 17a are the key regression diagnostics.

The hat matrix is most simply defined for the Normal theory model: $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$; $\bar{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$; $\hat{\underline{y}} = X\bar{\underline{\beta}} = X(X^TX)^{-1}X^T\underline{y} = H\underline{y}$. $H = X(X^TX)^{-1}X^T$ is said to put the "hat" on $\underline{y}$ (A rather fey nomenclature perhaps but surely no more than the one that invests elementary particles with "charm", "color", "strangeness", etc.). h_i is a measure of the distance of $\underline{x}_i^T$ from the centroid of the distribution of X (as we have seen). Values of $h_i > 2k/n$ identify "influential" observations. In the case of Poisson and probit regression models, the hat matrix diagonal, h_i, is a measure of both the weight of the ith observation described by the ith diagonal element of the weight matrix W and thus a function of the estimated response, $\bar{\mu}_i$, and of the distance (from the centroid) of $\underline{x}_i^T$. The abscissae, the index numbers, are simply the row numbers of the observations in $[\underline{y}, X]$. For this model of these data $2k/n = 2*3/19 = 0.32$. Therefore, those observations for which $i = 2,5,6$ and 16 are termed high leverage, or influential, observations.

The hat matrix diagonals can also be used to discriminate between interpolation and extrapolation for a regression model of a given set of data. For example, on the Normal theory model, an observation, say $\underline{x}_*^T$ for which $h_* = \underline{x}_*(X^TX)^{-1}\underline{x}_*^T > h_{(n)}$ where $h_{(n)}$ is the maximum valve of h_i for the model, is an extrapolation. For $h_* < h_{(n)}$ the observation is an interpolation.

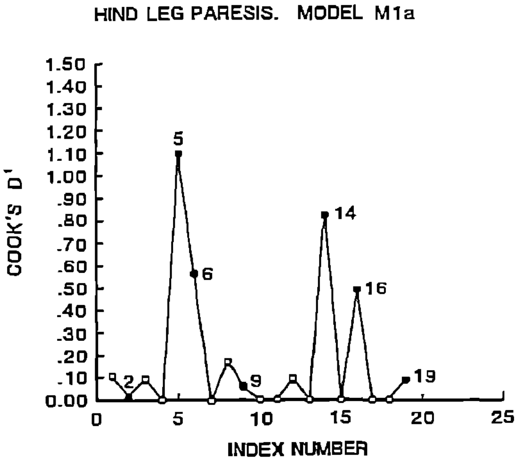


Fig. 18e is a case plot of D_i^1 for the empirical model M1a. D_i^1 is the one-step approximation to Cook's D_i for IRLS estimates of $\underline{\beta}$ for a generalized linear model. D_i^1, $1 \leq i \leq n$, is a measure of the change, $(\bar{\underline{\beta}}_{(i)} - \bar{\underline{\beta}})$ that occurs when the ith observation is deleted. $\bar{\underline{\beta}}_{(i)}$ is the estimate of $\underline{\beta}$ from the reduced sample (size n-1). The figure discloses that the estimates of $\underline{\beta}$ are dominated by these observations for which $0 < r_i < n_i$: #5, #6, #14, #16.

The extreme instability in the IRLS estimates of $\underline{\beta}$ for model M1a of these data is due to the joint effects of the several deficiencies in the experimental design remarked above. Note that Cook's D is a function of the key diagnostics e_i and h_i: $D_i^1 = k^{-1}[e_i^2/n_i\hat{\pi}_i(1-\hat{\pi}_i)]*[h_i/(1-h_i)^2]$ where $e_i = (r_i - n_i\bar{\pi}_i)$ is the ith residual and h_i is the ith diagonal element of the hat matrix for this model of these data. The hat matrix, H, is most simply defined for a Normal theory model: $H = X(X^TX)^{-1}X^T$. Since $\underline{y} = X\underline{\varepsilon} = X(X^TX)^{1}$-$Xy = Hy$, H puts the "hat", ¯, on y. h_i is a measure of the distance of $\underline{x}_i^T$ from the centroid of the distribution of X. There is a cognate definition for IRLS estimates, $\bar{\underline{\beta}}$.

In general, the regression diagnostics provide sensitive and detailed measures of the most important model-sample interactions (Herbert, 1986a). Belsley, et al (1980) have remarked that, "An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of the various estimates (coefficients, standard errors, ...) than is the case for most other observations. One obvious means for examining such an impact is to delete each row [of the observation matrix, [y, X]] one at a time, and note the resultant effect on the various calculated values. Rows whose deletion produces relatively large changes in the calculated values are deemed influential." For the Normal theory model a measure of the change, $(\bar{\underline{\beta}} - \bar{\underline{\beta}}_{(i)})$, in the OLS estimate, $\bar{\underline{\beta}}$, of the parameter vector, $\underline{\beta}$, that is produced by deletion of the ith row, $1 \leq i \leq n$, is Cook's standardized distance, $D_i = (\bar{\underline{\beta}}_{(i)} - \hat{\underline{\beta}})^T(X^TX)(\bar{\beta}_{(i)} - \bar{\underline{\beta}})/k\sigma^2$, where $\bar{\underline{\beta}}_{(i)}$ is the OLS estimate of $\underline{\beta}$ obtained from the sample, now size (n-1), with ith row deleted. For non-Normal theory models, Cook's distance is generalized to $D_i = (\bar{\underline{\beta}}_{(i)} - \bar{\underline{\beta}})^T(X^TWX)(\bar{\underline{\beta}}_{(i)} - \bar{\underline{\beta}})/k$. That is, the diagnostic for the influence on $\underline{\beta}$ of a given observation, $\underline{x}_i^T$, (row) is based upon perturbing the X matrix and evaluating the shift, $(\bar{\beta}_{(i)} - \bar{\beta})$, produced just as the diagnostic for influence on $\hat{\beta}$ of a given variable, X_i, (column) is based upon perturbing the correlation matrix $X_0^T X_0$, by adding a small multiple of the cognate identity matrix, and evaluating the shift, $(\bar{\beta} - \bar{\beta}^{**})$, in $\bar{\beta}$, produced thereby. $\bar{\beta}^{**}$ is the so-called Ridge regression estimate. See Figs. 35 and 36.) Values of $D_i = 1.0$ for a Normal theory model correspond to a shift of $\bar{\beta}_{(i)}$ to the 0.50 confidence ellipsoid on β. D_i is a function of the key diagnostics e_i and h_i: $D_i = [h_i e_i^2/\sigma^2(1-h_i)^2]/k$ where σ^2 is the mean residual sum of squares for the estimate $\bar{\beta}$. h_i is the ith diagonal element of the hat matrix for this model of these data.

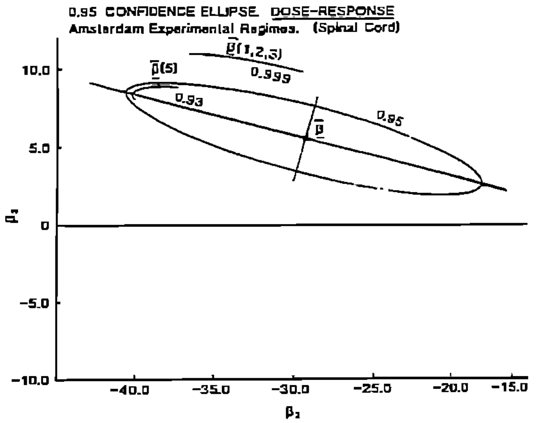


Fig. 18f is a plot of the $(1-\alpha) = 0.95$ confidence ellipse on β in the (β_2, β_3)-plane for the probit model M1a: $\eta = x^T\beta$. For a Normal theory model the ellipse is the locus of $(\bar{\beta} - \bar{\beta})^T(X^TX)(\bar{\beta} - \bar{\beta}) = F(\alpha; k, n-k)\sigma^2$, where $F(\alpha; k, n-k)$ is the α^{th} quantile of the F-distribution on k and n-k degrees of freedom. There are cognate expressions for Binomial and Poisson regression models. Two row-deleted estimates, $\bar{\beta}_{(i)}$ of β, are superimposed. It can be seen that $\bar{\beta}_{(5)}$, for which Cook's $D_5^1 = 1.10$, is shifted to the 0.93 confidence ellipse. Compare with Fig. 18e. Plotting the $\bar{\beta}_{(i)}$ in a space in which the metric is defined by $Var(\bar{\beta})$ provides an interpretation of D_i^1 that is absent from the index and variate plots of this diagnostic. Deletion of the 3 observations of the experiment for which N=1 gives the estimate $\bar{\beta}(1,2,3)$.

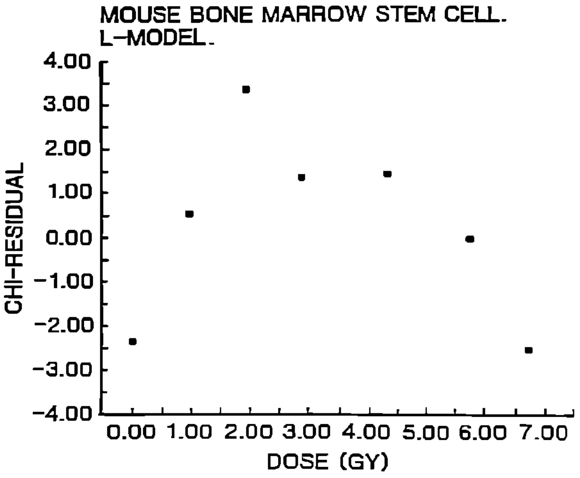


Fig. 19a. The plot is a scattergram of the chi-residuals, $\chi_i = (y_i - \hat{m}_i c_i)/\hat{m}_i c_i$, vs dose, for the linear, L, model $m_i = \exp(\beta_0 + \beta_2 D_i)$, of mouse bone marrow stem cell survival. There is a strong second-order pattern evident in the plot suggesting that the model should be augmented by a term D_i^2 to give the LQ model. The decrement, $\Delta\chi^2$ in the respective chi-squared statistics, $\chi^2 = \Sigma\chi_i^2$, for the L and LQ models is statistically significant. Note that these two models are nested and hence the decrement, $\Delta\chi^2$, is a valid discriminator.

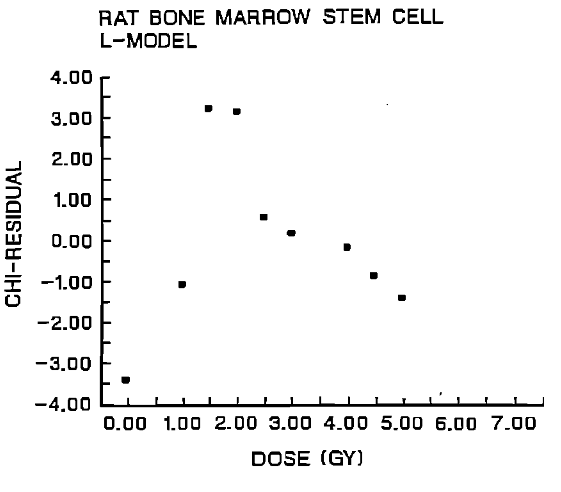


Fig. 19b. The plot is a scattergram of the chi-residuals, $\chi_i = (y_i - \hat{m}_i c_i)/\sqrt{\hat{m}_i c_i}$ vs dose, for the linear, L, model, $m_i = \exp(\beta_0 + \beta_1 D_i)$ of rat bone marrow stem cell survival. There is a strong pattern evident in the plot. Although the pattern is more complex than that of Fig. 19a it is quite clear that the fit would be improved by including higher order terms in dose, D, in the model. The decrement, $\Delta\chi^2$, in the respective chi-squared statistics for the nested L and LQ models is statistically significant.

Note that although there are only n=7 observations for the mouse data and only n=9 for the rat data, the respective residual analyses are still "meaningful", although the Snee (1977) criterion recommends at least n = k+10 = 13 observations for "... a meaningful residual analysis". This suggests that the conclusions from the residual plots of the LQ and T models of these data should be "meaningful", as well.

there are several statistical procedures to assist the investigator in discriminating between nested and non-nested rival models, for example, there are the AIC and PRESS statistics and the posterior odds ratio. These are discussed in sections 7.7 (AIC), 9.3 (PRESS), and 7.8 (posterior odds ratio). There is also the method of mixture, or "embedding", that is discussed in sections 7.1, 7.8, and 14.2. (See also Annex II, parts 3 and 5).

If a model does not fit the sample data then both the bias and variance of the sample estimates of parameters, $\underline{\hat{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$, as well as linear functions of parameters, such as the predicted response, $\underline{x}^T\underline{\hat{\beta}}$, and non-linear functions, $f(\underline{\hat{\beta}})$, such as $\hat{\theta} = \hat{\beta}_j/\hat{\beta}_k$ (e.g., the ratio α/β of the LQ models) will be inflated - depending upon whether the model is underfit or overfit, respectively. Deploying any model to convey the information contained in the data is a highly questionable procedure without first determining whether the model "fits" the data - on statistically adequate measures and criteria. For example, if the model is underfit because an important covariate, such as a measure of the duration of irradiation, say x_3 = log T, is omitted from the specification of the deterministic part of the model, μ_i, then the sample estimates of $\underline{\beta}$, etc. are biased $(E(\underline{\hat{\beta}}) \neq \underline{\hat{\beta}})$. See Annex II, parts 3 and 5 for examples of biased (aliased) estimates of $\underline{\beta}$. Figures 16-18 provide examples of the so-called category 2 model checks which "take a successful model to be one that leaves a patternless set of residuals (or other derived quantities)." Figures 18b and 18c suggest that the probit model M1a provides an adequate fit to the binomial data in the experiment described in Fig. 9; Figures 16 and 17 suggests that probit models M1b and M2(LQ) both underfit these data.

In the case of the multifraction LQ model examined in section 7.9.2 and in Annex II, part 3, residual analysis of the model, M1b, $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 = logD and x_2 = logN discloses that a time-factor, x_3 = log T, has been omitted that should have been included. Figures 16a, 16b, and 16c show that the linear predictor, $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N$, of the multifraction LQ model, M2, underfits the multifraction data of Fig. 9 to the same degree and in the same manner as does the linear predictor, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ of model M1b. However, Fig. 17 discloses that the fit of the latter can be markedly improved by adding the term x_3 = logT to give the model M1a with linear predictor $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Figure 18a shows that the fit of this model to these data is adequate.

However, based on the criterion of the distribution of the decrement in the residual sum of squares, ΔRSS, distributed as $\chi^2(1)$, the fit of the model M2(LQ) was not improved significantly by including a factor for time in the form γT to give the probit model M3 with linear predictor, $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3 T$. (Note that this is the generalized linear model appropriate to the so-called LQ + time hypothesis described in Fowler, 1989.) But it was found that the fit to the data of Fig. 9 was, on the same criterion, improved significantly by including the time factor in the form x_3 = logT, as was found to be the case for the model M1b. However, the resultant model, M4, with linear predictor, $\eta = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3 \log T$, was rejected by the data on the criterion of the Pearson chi-squared statistic, $RSS = \chi_c^2 = \Sigma\chi_i^2$. As we remarked earlier, model selection based only on decrements of chi-squared (or deviance), in the absence of a check on the overall fit, may not lead to a useful model since the best-fitting of two rival models may not fit very well. The fact that model M3 does not fit these data, suggests that either the time factor proposed by Travis and Tucker (1983) namely, -γT, or the LQ parameterization of the factors for D and N - or both - should be re-examined. (N.B. It is shown in Annex II, part 3 that the probit model M5a with linear predictor, $\eta = \beta_0 + \beta_1 D + \beta_2 D/N + \beta_3$ T, fit the data quite well: $\chi_c^2 = 21.2$, $P(\chi^2 > \chi_c^2|15) = 0.13$, $\hat{\beta}_j/\sqrt{Var(\bar{\beta}_j)} > 6.0$, $0 \leq j \leq 3$. The model M5a is, of course, also a Taylor series model.) The above evidence that the addition of a time factor, logT, markedly improved the goodness-of-fit of all models of the Tucker and Thames (1983) data (obtained from van der Kogel, 1979) on spinal cord injury (Fig. 9) suggests that there is a strong time dependence to late effects. This refutes the statement by Fowler (1989): "It should be stressed that no time factor is required for late effects ..."

7.9.2 Will the LQ model "fit" cell-survival data? Did it ever fit the BEIR III LSS leukemogenesis data?

"The use of the LQ model [of cell-survival] was dictated solely by its outstanding descriptive properties."

B. Fertil and E. Malaise, 1985

"Although the differences among models [of leukemia incidence] with respect to goodness-of-fit are not large, they suggest dependence on both gamma dose and its square."

BEIR III Report, 1980

"These are the two extremes that science must avoid: theories that don't agree with reality, and theories that can be made to agree with anything."

I. Stewart, 1992

We now examine two cases in which the linear quadratic (LQ) hypothesis leads to model misspecification; in one case it underfits, in the second case it overfits the data on which it is deployed.

Figures 19a and 19b are added-variable plots of the chi-residuals χ_i vs dose D_i, for the linear Poisson model, $m_i = \exp(\beta_0 + \beta_1 D_i)$, $1 \leq i \leq n$, for the rat and mouse stem cell survival data reported in Frome and Beauchamp (1969) and Till and McCullough (1960), respectively. Here m_i is the Poisson rate constant. Both suggest that the respective linear Poisson models underfit the cell survival data. Both plots are predominantly of second degree in the dose D suggesting that a term D^2 should be added to the model, although the plot for the rat data suggests the possible presence of still higher order terms in D.

Figures 20a and 20b are the added-variable plots of the chi-residuals, χ_i vs dose D_i, for the Poisson model of the LQ hypothesis, $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$, $1 \leq i \leq n$, for the mouse and rat stem survival data of Fig. 19. In both plots the largest residuals are in the dose range $0 \leq D_i \leq 2.0$ Gy, showing that the LQ model does not fit very well in the so-called "shoulder region", contrary to the received beliefs on this issue (See Fertil et al, 1983 and Fowler, 1984). The plot of Fig. 20a is weakly fourth degree in the dose D. However, the plot of Fig. 20b has a strong third degree pattern suggesting that a term D^3 should be added to the LQ model. Taken together, Figs. 19 and 20 suggest that the LQ model of cell-survival is simply a Taylor series approximation (to terms of second order) to the dose-response curve of the "true" cell survival model whatever that may be (we shall see below why it might be the Target theory model). It also appears from Fig. 20b that the LQ model may be a very poor approximation in some cases. Figures 20c and 20d, which are plots of Cook's distance vs dose for the rat and mouse cell survival data, respectively, show that the parameter estimates of LQ models may be exceedingly labile: The estimates are not invariant under row-deletion; that is, $\underline{\hat{\beta}}_{(i)} \neq \underline{\hat{\beta}}$, $1 \leq i \leq n$, where, it will be recalled, $\underline{\hat{\beta}}_{(i)}$ is the estimate of $\underline{\beta}$ obtained from a set of the original data from which the i^{th} observation has been deleted. But, as Box et al, 1978 have remarked, "In an adequate model, constants stay constant when variables are varied", a remark that under-scores the inadequacy of the LQ model for these data. One reason for this excessive lability of the sample estimate of $\underline{\beta}$ for the LQ model is that unless the range of dose is quite large, the variables D_i and D_i^2 will be highly correlated and hence the matrix X^TX will be nearly singular; it will be recalled that, for the Normal theory model, $\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}} = -(X^TX)^{-1}\underline{x}_i^T e_i/(1-h_i)$. A similar dependence of $(\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})$ on $(X^TWX)^{-1}$ obtains for models of Binomial and Poisson responses.

The evidence of Figs. 20a and 20b suggests that the single fraction LQ model represents only the first two terms of a Taylor series approximation to the correct, mechanistic, model of the observed response, cell-survival, and that for that response in some species, a second degree approximation will underfit the data. But it will be recalled that the received form of the multifraction LQ model is derived by cascading the single-fraction forms: $S_N = \Pi S_1$ where $S_1 = \exp(\alpha d + \beta d2)$. Clearly, if S_1 is underfit then it is quite likely that S_N will be underfit also. In general, if a model is underfit then the sample estimates of the parameter vector $\underline{\beta}$ are biased (aliased) by amounts that depend on both the functional relationship of the omitted variable(s) to the response and on the distribution of observations in the sample. And again, if the sample

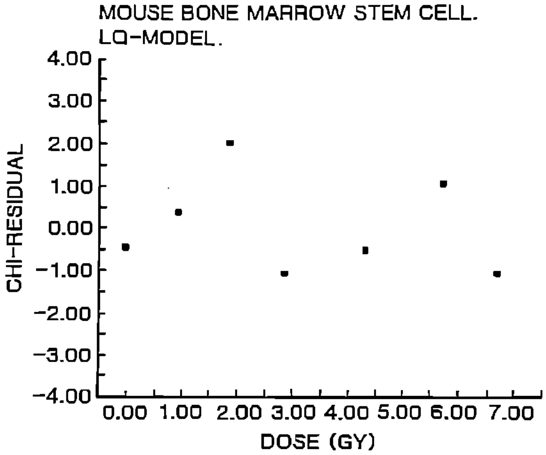


Fig. 20a. The plot is a scattergram of the chi-residuals, χ_i, vs dose, D_i, for the LQ model, $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$ of the mouse bone marrow stem cell survival data. Although there is some evidence of serial correlation in the plot suggesting the presence of a week fourth order pattern, all but one of the residuals is less than 2.0. Together with the evidence of the aggregate statistic, $\chi_c^2 = \Sigma\chi_i^2 = 8.138$, $P(\chi^2 > \chi_c^2|4) = 0.09$, the plot suggests that the LQ model provides an adequate fit to these data. Note however, that the poorest fit - largest residual - is in the so-called "shoulder-region" $1.0 \le D_i \le\!<2.0$ Gy.

However, the LQ model does not fit as well as the rival Target (T) model, which, for the same number of adjustable parameters (k=3) gives $\chi_c = 7.597$, $P(\chi^2 > \chi_c^2|4) = 0.11$. It is also the case that, on the criterion of predictive performance in new data, the Target theory model is superior. The respective PRESS statistics for the Target and LQ models are 12.0 and 13.3. Moreover, it will be seen below that the rival Target model fits better than the LQ model in the "shoulder region".

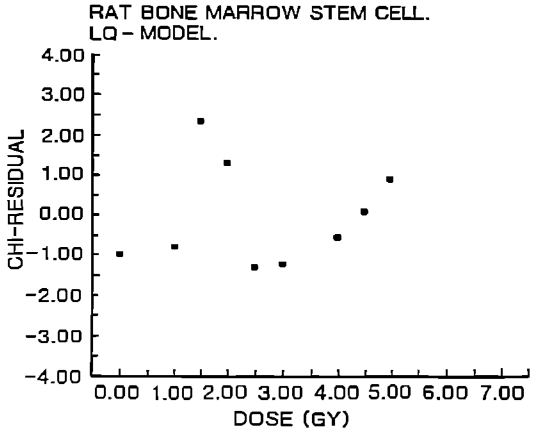


Fig. 20b. The plot is a scattergram of the chi-residuals, χ_i, vs dose, D_i, for the LQ model of the rat bone marrow stem cell survival data. Note that, as in the case of the LQ model of the mouse data, the largest residuals = $\chi_i > 2.0$ - are in the shoulder-region, $1.0 < D_i < 2.0$ Gy. (N.B.: "Malaise et al ... also found that the Linear Quadratic model is best suited to fitting the initial part of the cell survival curves ..." R. Yaes, 1987.) There is strong serial correlation present and a definite third-order "pattern", suggesting that the model should be augmented by a term in D_i^3; that is, the LQ model underfits these data, leading to biased (aliased) estimates of the parameter vector β. (Underfitting results from an excess of parsimony.) This evidence of the case statistics, χ_i, is consistent with that of the aggregate statistic $\chi^2 = \Sigma\chi_i^2 = 13.214$, $P(\chi^2 > \chi_c^2|4) = 0.04$; these data reject the LQ model. One possible alternative is the LC model: $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i^3)$. The decrement, $\Delta\chi^2$, in the respective chi-squared statistics, $\chi^2 = \Sigma\chi_i^2$, for the LQ and LC models is statistically significant. This was also the case for the rival LC* model, $m_i = \exp(\beta_0 + \beta_2 D_i^2 + \beta_3 D_i^3)$, which provided the better fit. It gave a smaller value of χ^2. Moreover, it is the more parsimonious. However, for the LC model the slope of the survival curve at D=0 is negative, while for the LC* model it is zero. The LC model is consistent with this aspect of the received wisdom on the matter of slope; the LC* model is not. As Robins and Greenland (1986) have remarked, one should "... choose a model that is consistent with the data and yields parameter estimates consistent with ... prior beliefs." But surely, if that model does not fit the data as well as a rival model for which the parameter estimates are inconsistent with prior beliefs, then one should re-examine the prior beliefs rather carefully.

Note that in this case, the LQ model was rejected on the evidence of the goodness-of-fit measures, RSS, to present data, whereas, in the case of the mouse data, the LQ model was rejected on the evidence of the goodness-of-fit measure, PRESS, to future data.

It is apparently believed by some eminent investigators that the precision of the parameter estimates of a model provides an adequate measure of the concordance of the model and data: "… we obtained a survival curve characterized by the fitted parameters. The experimental fluctuations and the quality of the fit [emphasis added] were expressed by both the variances and covariance(s) linked to these parameters. … the experimental fluctuations and the quality of fitting [emphasis added] are represented by a 95% confidence ellipse or ellipsoid" (Fertil, Deschavanne, Lachet, and Malaise, 1980). However, this is surely not the case. The proper measure of the "experimental fluctuations" and the "quality of the fit" is the sum of squared residuals, $RSS = \Sigma e_i^{*2}$, together with the several plots of the residuals, e.g., e_i^* vs D_i, etc. In the present case - the LQ model of the rat data - the precision of the parameter estimates is adequate, $\bar{\beta}_j/\sqrt{Var(\bar{\beta}_j)} > 3.5$, $j = 1,2$, but it is obvious that the LQ model does not "fit" these data.

N.B.: It can be shown that the bias in the sample estimate $\bar{\beta}$ for the LQ model of the mouse data is a function of both β_3 and the design matrix X. That is, $E(\bar{\beta}) = \beta + A\beta_3$ where A is the so-called alias matrix: $A = (X^TWX)^{-1}XWX_4$, where X is the 9*3 design matrix for the LQ model of the rat data. X_4 is the 9*1 matrix with elements D_i^3, $1 \leq i \leq 9$. That is, "The bias terms … depend not only on the postulated and true models but also on the experimental design …" (Draper and Smith, 1981).

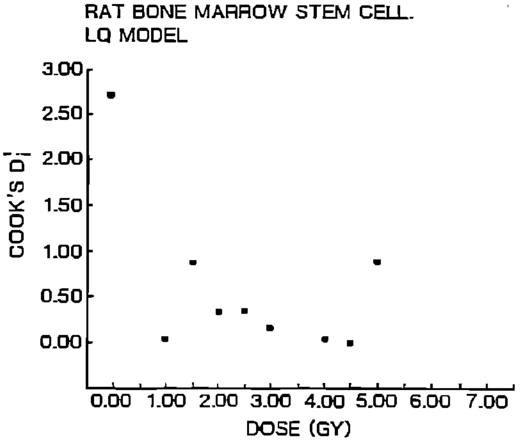


Fig. 20c. The plot presents a scattergram of D_i^1 vs dose D_i for the LQ model of the rat bone marrow stem cell survival data where D_i^1 is Pregibon's one-step approximation to Cook's D_i for non-Normal (Poisson) data. The cut-off of D_i^1 may be taken to be ~ 0.2. It is evident that the sample estimate of β for the model is dominated by the observation at $D_1 = 0$.

The quadratic function is quite "rigid", a property which makes it a useful tool for "smoothing" data - as in the method of central differences - in order to reduce the effects of the random component in the observed response. However, this property of rigidity makes the estimates of the parameter vector of the quadratic function very sensitive to the presence of single observations.

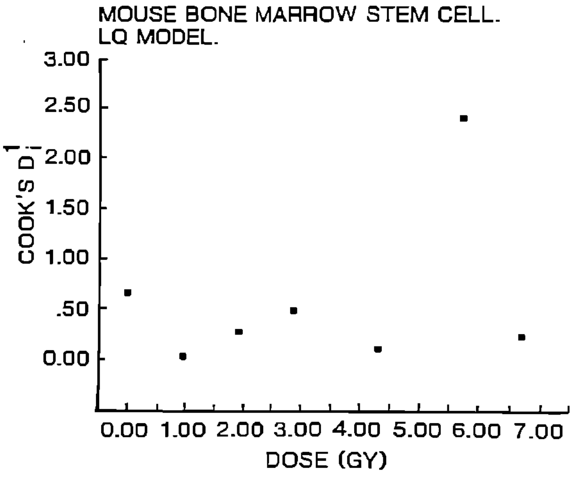


Fig. 20d. The plot presents a scattergram of D_i^1 vs dose D_i for the LQ model of the mouse bone marrow stem cell survival data. It is evident that the sample estimate of β is dominated by the observation at $D_6 = 5.70$ Gy. It is evident that the parameter estimate, $\bar{\beta}$, of the LQ model of the mouse data is less labile than for the rat data. This may be due to the decreased correlation between D and D^2 in the mouse data since the range of dose is greater: 6.70 Gy vs 5.00 Gy (Recall that $\beta_{(i)} = \beta - (X^TX)^{-1} e_i x_i/(1-h_i)$ for the Normal theory model).

estimates of $\underline{\beta}$ are biased then, of course, so are functions of $\underline{\beta}$ such as the response and ratios such as $\theta = \alpha/\beta$.

On the other hand, if the model is overfit, because a redundant variable, such as D^2, is included in the specification, then the variances, $Var(\hat{\underline{\beta}})$ and $Var(\underline{x}^T\hat{\underline{\beta}})$, of the sample estimates of the parameter vector and response, respectively, are inflated, and the bias as well as the variance of estimates of non-linear functions, $\theta = f(\underline{\beta})$ of the parameter, say the ratio of two regression coefficients, $\theta = \beta_j/\beta_k$, is also inflated. The Poisson models of the LSS data on leukemia incidence used for the BEIR III report (1980) provide a vivid example of the effects of over-fitting a set of data and hence of the utility of Ockham's principle of parsimony. We examine this matter next.

In some studies the aggregate statistic of concordance, say χ^2, (or decrement thereof) is either under- or over-interpreted so that the a priori evidence for the model is allowed to dominate the sample evidence, which is summarized by the aggregate statistic, against it (a motivationally-based error). For example, it is obvious in Table V-8 of the BEIR III Report that the Committees' "model of choice", the LQ-L, based on $\Sigma\chi_i^2 = 10.4$ on 11 df, overfits the BEIR III LSS leukemia incidence data since the decrement in $\Sigma\chi_i^2$ between the LQ-L model and either rival (L-L or Q-L) is not significant. (The decrement is less than 3.84, the 0.95 quantile of Pearson's chi-squared distribution on 1 degree of freedom.) It will be recalled that in the BEIR III Report, the Committee also chose the L-L model of female breast cancer incidence in the LSS sample despite the fact that it can be readily shown that the model overfits the data, since the estimate of the coefficient of the neutron dose, D_n, is only somewhat over half as large as its standard error. But situations in which the a priori evidence on an issue dominates the sample evidence should surely only rarely occur in any investigation in empirical science (as Box and Tiao (1973), for instance, have pointed out).

Figure 21a is a so-called added-variable plot of the chi-residuals for the BEIR III L-L model of leukemia incidence. Note that, unlike the cognate plots for the L-model of cell-survival in Figs. 19a and 19b, there is no evidence of a second-order pattern that would suggest that a term D_γ^2 be added. (D_γ denotes the gamma dose.) Thus, the evidence of this (Category 2) model check is consistent with that based on the decrement in the respective chi-squared statistics presented in Table V-8 in the BEIR III report (NAS/NRC, 1980): The LQ-L model overfits the LSS leukemia incidence data. (Moreover, in Annex III, part 4 it is shown that the observation at $(D_\gamma, D_n) = (38.8, 0.1)$, index number 12 in Figs. 21b and 21c, provides most of the weak empirical evidence for the LQ-L model vs the L-L model.) It may be expected that the bias of the LSS sample estimate of the cross-over dose, $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$, will be inflated. As indeed we have found it to be: The maximum likelihood estimate is $\hat{\theta} = 117$ rad; the reduced-bias (weighted jackknife) estimate is $\hat{\theta}_J = 17$ rad. See Hinkley, 1977 and section 7.3 and Annex III, part 4. (N.B. Although it is obvious from Table V-8 of the BEIR III report that the LQ-L model overfits the LSS data, nonetheless the LQ-L was the "model of choice" in that report - as we have remarked elsewhere in this report.)

In the case of the single-fraction LQ model of cell-survival, residual analysis discloses that it underfits the data; i.e., that additional cubic or fourth order terms in dose are required to "fit" the data. In the case of the BEIR III leukemogenesis data the single fraction LQ model overfits the data. These findings suggest that the LQ model is not a mechanistic model but simply a Taylor series approximation to the correct, mechanistic, model, in each case.

Figures 24a - 24f provide plots of the regression diagnostics for the LQ-L model of the BEIR III leukemia incidence data. These plots suggest that even as a Taylor series approximation, this model is altogether inadequate. For example, it is clear that there are several observations that are not well "approximated" by the model, and several others that dominate the sample estimates of the model parameter vector, $\underline{\beta}$.

However, it must be acknowledged that although the LQ model does not fit the BEIR III leukemia incidence data according to the best statistical measures and criteria, it has been shown that a model with a linear-quadratic dependence on dose, on these criteria, fit the BEIR V 1990 leukemia mortality data. However, there are several important distinctions between these two sets of data - besides that between incidence and mortality. These are discussed at length in the BEIR V report (NAS/NRC, 1990). We mention only three here: 1) the DS86 (BEIR V) dosimetry is "better" (more accurate, more precise) than the T65D (BEIR III) dosimetry. 2) The BEIR III data are augmented (n = 16); the BEIR V data are disaggregated (n = 2575). The respective decrements of deviance, ΔD, between the linear and linear-quadratic models are $\Delta D = 1.97$ (BEIR III) and $\Delta D = 4.00$ (BEIR V). (Herbert, 1990). For both comparisons ΔD is distributed as Pearson's chi-squared on 1 df, $\chi^2(1)$. It will be recalled that the 0.95 quantile of $\chi^2(1)$ is 3.84. The Freeman-Tukey residuals, g_i, for the linear-quadratic models include one outlier ($g_i > 2.0$) for the BEIR III data and none for the BEIR V data. However, the BEIR III report preferred the LQ-L model for the non-leukemia mortality data as well as the leukemia incidence data. Moreover, the report required that ratios of the parameters of the LQ-L model of non-leukemia mortality data agree with cognate ratios obtained from the sample estimates of the LQ-L model of the leukemia incidence data in order to stabilize the sample estimates of the parameters of the LQ-L model of the non-leukemia mortality data. That is, the ratios of the LSS sample estimates of the parameters of the LQ-L model of the leukemia incidence data were imposed as a priori constraints in obtaining the (constrained) estimates of the parameters of the so-called LQ-L model of non-leukemia cancer mortality in Table V-11 of the BEIR III (1980) report. But it has been shown in the BEIR V report that the linear-quadratic model is only valid for leukemia mortality, not for non-leukemia mortality. This matter will be discussed further below (See section 14.3.) It should also be noted that the BEIR V leukemia mortality data are much weaker than the BEIR V non-leukemia mortality data. For example, although there are 2575 records for leukemia mortality there are only 173 leukemia cases, whereas for digestive tumor mortality there are 2162 records and 2193 cases. Thus, the evidence for the model that is linear in dose is stronger than the evidence for the model that is linear-quadratic in dose (NAS/NRC, 1990).

The evidence of Fig.16-18 suggested that the multifraction LQ model may be mis-specified for some kinds of data because it is underfit - a time factor is omitted. However, the evidence of Figs. 19-20 suggests the presence of a more fundamental weakness: The form of the dependence of the fraction of surviving cells, S_i, for each fraction of dose, D_i, $1 \leq i \leq N$, may be also mis-specified. In this event, the form of the multifraction LQ dose response model of survival, S_N,

$$S_N = \overset{N}{\Pi} S_i$$

at N equal fractions of dose, is incorrect, even if one grants that the decremental survival, S_i and D_i, is independent of the position of D_i, $1 \leq i \leq N$, in the sequence of N fractions. It is therefore, especially unfortunate that the method by which the form of the multifraction, $N \geq 1$, LQ model was derived - the cascading of N single-fraction LQ models of cell survival - confers on it something of the ontological status and properties - especially that of valid extrapolations - of a mechanistic model. The method of its derivation - from a model of a more elementary effect - implies that the multifraction LQ model must be "law-like". (Many of the better known of the radiobiological modelers were first trained in physics - "the ultimate reductionist subject" (P. Anderson, 1991). Hence their models often describe radiation effects in tissues in terms of models of radiation effects in cells.) The multifraction LQ model is an example of what I.D.J. Bross (1970) has described as a deep model: "The goal of a 'deep' model is to predict the observable events from the parameters that characterize the events at the microscopic level." Figs. 16-20 suggest that this may not be the case at all, but rather that the multifraction LQ model may instead simply represent an often rather poor approximation to the process that generated the data. Thus, it cannot be assumed, a priori, that it will provide an adequate "fit" to any sample of data in which it may be deployed. That is, the concordance of the multifraction LQ model and data must be checked - by statistically adequate methods - in every case, especially in extrapolation.

The secondary analyses presented, in sections 7.9.1 and 7.9.2, of the data and models provided in three published studies on cell survival, carcinogenesis, and hind-leg paresis in rodents strongly suggests that the general LQ hypothesis suffers from both of the "extremes" that I. Stewart has warned against (vide supra): It (apparently) does not, "agree with reality" and it (apparently)

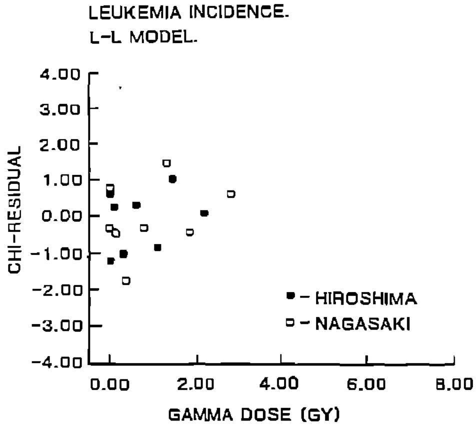


Figure 21a. The plot is a scattergram of the chi-residuals, χ_i, vs dose, D_γ, for the L-L model of the BEIR III leukemia incidence data. The L-L model is, $m_i = \beta_0 + \beta_1 D\gamma + \beta_2 D_n + \beta_3 C$, where D_γ and D_n are, respectively, the gamma and neutron doses, and C is a (0,1) indicator variable that identifies the residence, Hiroshima or Nagasaki, of those irradiated. The filled symbols identify observations at Hiroshima. The open symbols identify the observations at Nagasaki. The distribution is essentially the "white noise sequence" required by Box, Hunter and Hunter, 1978 for a "statistically adequate model"; it is unrelated to the variable D_γ . In particular, there is no evidence of any pattern in the residuals that would suggest that the model be augmented by including a term D_γ^2. That is to say, the L-L model appears to have retrieved all of the information on dose-response that is included in these data - none has "leaked" into the residuals. Compare with Figs. 19a and 19b.

This evidence of residuals is consistent with that of Table V-8 of the BEIR III Report in which it is evident that the decrement, $\Delta\chi^2$, is the respective chi-squared statistics, $\chi^2 = \Sigma\chi_i^2$ of the nested L-L and LQ-L models is not statistically significant: $\Delta\chi^2 = 1.10$, but the 0.95 quantile of the distribution of χ^2 on 1 df is 3.84. Compare Fig. 21 with Fig. 19.

It is clear from Fig. 20b that the LQ model underfits the rat bone marrow stem cell survival data - a term in D^3 must be included - and hence the estimates of β_1 and β_2 are biased (aliased. vide supra). However, it is equally evident from Fig. 21a that the LQ-L model over-fits the leukemia incidence data - the quadratic term D_γ^2 is redundant and simply serves to inflate the covariance matrix, $Var(\bar{\beta})$, and therefore inflates the bias - as well as the variance, $Var(\theta)$ - of the sample estimate, $\bar{\theta} = \bar{\beta}_1/\hat{\beta}_2$, of the cross-over dose $\theta = \beta_1/\beta_2$, since it is a non-linear function of β. (As F.S. Acton (1959) has remarked: "Some data fit lines naturally, but others have lines thrust upon them.")

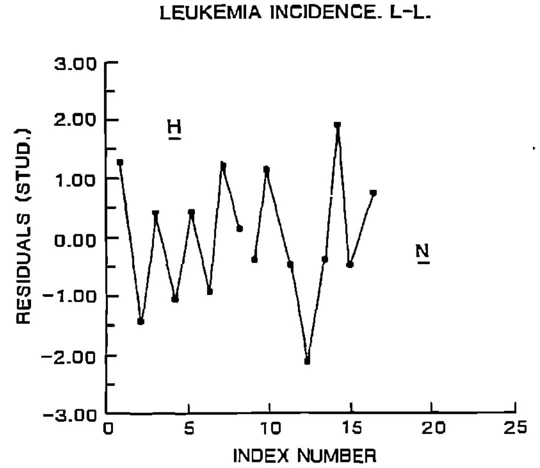


Fig. 21b. The figure presents an index plot of the Studentized residuals for the L-L model of the BEIR III leukemia incidence data.

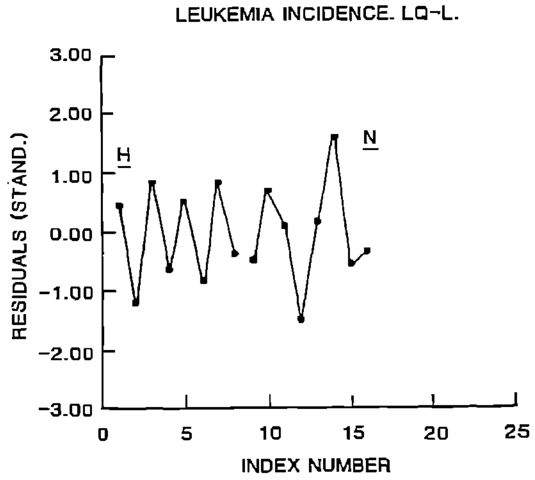


Fig. 21c. The figure presents an index plot of the Studentized residuals for the LQ-L model of the BEIR III leukemia incidence data. Comparison with the cognate plot of Fig. 21b discloses that the addition of the term D_γ^2 to the L-L model to give the LQ-L model does not materially improve the fit.

Figures 21a - 21c show quite vividly that there is no empirical evidence to support the LQ-L model of the BEIR III leukemia incidence data.

Robins and Greenland (1986) have remarked that, "... both improper omission and indiscriminate inclusion of variables in a model can lead to compromised inferences." Figures 20 (The LQ model underfits) and 21 (The LQ-L model overfits), respectively, provide examples of each.

has been "made to agree with anything". (The topologist Rene Thom has remarked that, "Theories that explain everything, explain nothing.")

7.9.3 <u>The lability of sample estimates of the α/β ratio.</u>

"I shall use the $\alpha d + \beta d^2$ model because I believe it to be ... valid for doses per fraction in the radiotherapy range, i.e., up to about 10 Gy ..."

J. Fowler, 1984

We have shown above that the sample estimates of the α/β ratio of the LQ model may well be inherently meaningless and that even if it had meaning, the sample estimate, $\hat{\theta} = \hat{\alpha}/\hat{\beta}$, may be biased. We now show that the <u>reported</u> sample estimates of α/β may be also quite <u>labile</u>, in the sense that the presence or absence of a single observation in the sample may profoundly alter that estimate.

In some of the F_e-plots constructed for LQ models that have been reported in the peer-reviewed literature, the sample estimate of α/β that is obtained is <u>dominated</u> by an extreme, or remote, observation at N=1, D/N $\doteq$ 15 Gy which lies well beyond the stipulated range of validity, $0 \leq D/N \leq 10$ Gy, of the multifraction LQ model. We found that if this observation is deleted, the estimate of this ratio may be altered by a factor of as much as two. We have remarked in Fig. 2a that this effect may result, in part, from the marked non-uniformity in the distribution of observations in the design of the experiment from which the data were generated, a generic weakness that seems to appear, in varying degree, in many of such experiments that are reported in the literature (see Figs. 2a, 10, and 22). It will be useful (both to emphasize the effect and to illustrate the simple case statistics required) to describe this weakness - and its effects - more concisely and vividly in terms of the key diagnostics, e_i^*, the Studentized residual (RSTUDENT) and h_i, the hat matrix diagonal, for the model $D^{-1} = (\alpha/\log S) + (\beta/\log S)D/N$ for the data of the study shown in Fig. 22, a Normal -theory model. The probability plot of the residuals, e_i^* is shown in Fig. 23a and the scattergram of the e_i^* vs h_i in Fig. 23b. It is evident that the observation at d = D/N = 15 Gy may <u>dominate</u> the sample estimates of α, β, and α/β since the corresponding hat matrix diagonal, h_8, is nearly twice the cut-off 2k/n = 0.44. The index plot in Fig. 23c shows that the observation at d = 15 Gy does, in fact, dominate the sample estimates of α/β since the sample estimate of $\theta = \alpha/\beta$ $(=\beta_0/\beta_1)$ changes by about a factor of two upon deletion of this observation. (The row-deleted estimates, $\hat{\theta}_{(i)}$, described in Fig. 23c are obtained by the mean-shift outlier method. See section 7.1.) Obviously, such labile estimates should be used with great caution - if at all. Such lability warns that the data is <u>inadequate</u> to provide estimates of either model parameters or of functions of model parameters, such as the response - in either interpolation or extrapolation - and, especially, such estimates are inadequate for generalization to other data sets. See Annex III, part 3. (Figure 23 describes a "finger exercise" in diagnostics.)

We have remarked in section 7.5 and in Annex II, part 3, that in a sample of observations on a binary response to multifraction, $N \geq 1$, radiation treatment regimens, the target tissues for N=1 must differ <u>qualitatively</u> from those for N > 1 since the former had not been previously irradiated and thus the increments in response at each increment of dose in the two regimens may differ in <u>level</u> and in <u>kind</u>. Thus, the sample is, a priori, <u>heterogeneous</u>. The sensitivity of the sample estimates of α/β that are obtained from F_e-plots of such data to <u>deletion</u> of the observation at N=1 provides sample evidence of this heterogeneity: The estimates of α/β obtained from the F_e-plots in Figs. 2a and 22 both change by a factor of about two when the observation at N=1 is <u>deleted</u>. See Annex III, part 3. (Perhaps a useful physical metaphor can be found in the difference between the <u>transient</u> and <u>steady state</u> behaviours of a periodically driven oscillator.)

Let us examine the issue of the lability of the sample estimates of the ratio, θ, for an LQ model in still another example. Consider, for instance, the Poisson regression models. Take the LQ-L model of the BEIR III leukemia incidence data. Figure 24a presents a plot of the observed, y_i, vs expected, $\hat{\mu}_i$, levels of response for this model of these data. Figure 24b presents a Normal probability plot of the Freeman-Tukey residuals, g_i. On the evidence of the Filliben probability plot correlation coefficient, $r_F = 0.978$, a Normal distribution of residuals cannot be rejected. A

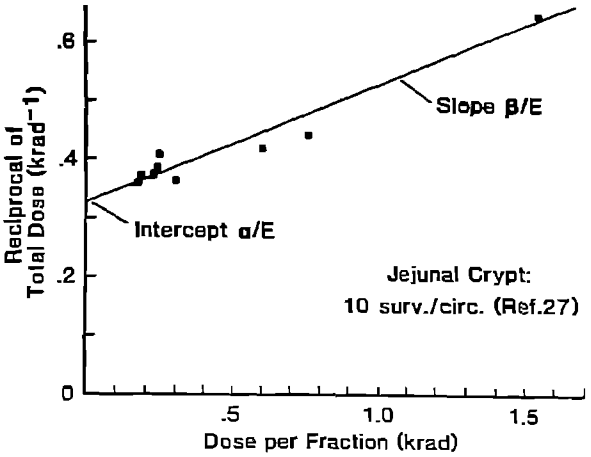


Fig. 22. The plot is a scattergram of $D^{-1}(S)$ vs $D(S)/N$ for a specified level of survival, S. (Reproduced with permission from Thames et al, 1982.) The F_e-plot of these data is super-imposed. Note first that the distribution is quite non-uniform: 8/9 or 89% of the observations lie in the lowest 50% of the range of D(S)/N. Note next that the extreme observation, $D(S)/N \simeq 15$ Gy, that will dominate the sample estimate of α/β for which the F_e-plot is constructed to obtain, lies well beyond the stipulated upper limit of dose per fraction, 10 Gy, for which the LQ model is valid (Fowler, 1984). The sample estimate of α/β that is obtained with this observation included is $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2 = 16.81$. If this observation is deleted this estimate is $\bar{\theta} = \hat{\beta}_1/\bar{\theta}_2 = 30.14$; the estimate is inflated by a factor of 1.79. Note also that observations for which N=1 are qualitatively distinct from those for which $N \geq 2$ since in the former case the observed response is for previously unirradiated tissue, while in the latter case the response is that for previously irradiated tissue. Hence, data that include both kinds of observations are hetero-geneous. (N.B.: "While there may be no simple way to obtain the individual LQ model parameters for these cells, the α/β ratio can be obtained from the Fe isoeffect plot of Douglas and Fowler" R. Yaes, 1988. The quotation cites Thames et al, 1982.)

plot of the g_i vs $\hat{\mu}_i$ is presented in Fig. 24c. The residual $g_{12} = -2.16$ at $y_{12} = 0$ lies outside the usual "cut-offs" at $g_i = \pm 2.0$. The evidence of the three plots suggests that the LQ-L model "fits" the data; however, the response $y_{12} = 0$ is not "well-explained" by the model. Note that it is shown in Annex III, part 4 that most of the (admittedly weak) empirical evidence for the LQ-L model of leukemia incidence presented in the BEIR III Report resides in this single anomalous (Nagasaki) observation: a zero level of response at a dose well above zero: $(D_\gamma, D_n) = (38.8, 0.1)$ - dose in rad. If this observation is deleted then the empirical evidence for the LQ-L model, already weak, disappears altogether from the BEIR III data. And in the absence of this observation the sample estimate of the cross-over dose, $\theta = \beta_1/\beta_2$, changes from 117 rad to 362 rad (see Annex III, part 4). However, unlike the case of the F_e-plots cited above, this influential observation is not the most remote observation of the sample; there are several at larger values of dose. But note that, on the evidence of the Freeman-Tukey residual, it is distinguished as an "outlier": $g_{12} > 2$.

Figures 24d-24f are index plots of DFBETAS for the LQ-L model. They are included here to vividly demonstrate the value of examining, in addition to the residuals, still other case statistics of models that, on the evidence of aggregate statistics, such as χ^2, and plots of its components, the residuals, χ_i are apparently "adequate". The diagnostics described in Figs. 24d-24f were estimated from the form, $P\underline{y} = PX\underline{\beta} + \underline{e}$, where $PP^T = V^{-1} = W$ the (n*n) Poisson weight matrix obtained at the final iteration of the iterative reweighted least squares procedure by which the sample estimate of $\underline{\beta}$ was obtained (See Belsley et al, 1980; Theil, 1971).

Figure 25a, an index plot of the row-deleted estimates, $\hat{\theta}_{(i)} = \hat{\beta}_{1(i)}/\hat{\beta}_{2(i)}$, $1 \leq i \leq n$ (=16) for the LQ-L model of the BEIR III leukemia incidence data shows the influence of this single observation on the sample estimate of the so-called cross-over dose, $\theta = \beta_1/\beta_2$, quite clearly. (N.B. It should be recalled that in the NIH Radioepidemological Tables (1985) the value of the cross-over dose obtained from the BEIR III leukemia incidence data is given as $\hat{\theta} = 117$ rad.)

Figure 25b discloses that the sample estimate of the ratio $\beta_1/\beta_2 = \alpha/\beta$ for the LQ model of the rat cell survival data discussed earlier is dominated by the single observation at $D_1 = 0$. Figure 25c discloses that the sample estimate of the ratio $\beta_1/\beta_2 = \alpha/\beta$ for the LQ model of the mouse cell survival data discussed above is labile but is not dominated by a single observation. It should be recalled that the LQ model "fit" the mouse cell survival data much better than it did the rat cell survival data (The rat data rejected the LQ model on the evidence of both aggregate and case statistics. See Fig. 20b).

The row-deleted estimates, $\hat{\theta}_{(i)} = \hat{\beta}_{1(i)}/\hat{\beta}_{2(i)}$, described in Figs. 25a-25c, are obtained by the mean-shift outlier method (see section 7.1).

Note that the presence of such weaknesses in estimates and inferences as discussed above, which arise as consequences of either the model selection (misspecification of the deterministic and/or random parts of the response) or the sample data (presence of outlying and influential observations and/or collinear variables) and their respective effects on study findings (estimates and inferences) are disclosed at once to an examination of regression diagnostics (the so-called category 1-3 model checks), including also the PRESS statistic, but not necessarily to a goodness-of-fit test based on the sampling distribution of an aggregate statistic such as $\Sigma\chi_i^2$ or Σd_i^2. For, as Robins and Greenland (1986) remark, "... a problem with goodness-of-fit tests is that they are insensitive to certain types of inconsistencies between models and data, so that they can indicate a good fit for certain models that are grossly inconsistent with the data. Thus, if one wishes to use a model consistent with the data, one ought to check for model inadequacy by examining residuals, screening for outliers, and employing other similar 'diagnostic techniques'." But even here judgment is required; as Nelder (1968) observes, "A question that has occupied the attention of modelers is that of how well some specified model fits the data ... The methods usually involve examination of the residuals, the e_t, that being the differences between the observed responses Y_t and the fitted or predicted responses. Daniel and Wood (1972, App. 3A) show cumulative distribution plots of generated random normal deviates with mean zero and unit variance. ... Although sets of 32 are visibly better behaved, they still exhibit patterns which could easily be assumed by the modeler to be departures from randomness even though the data are known to be random. Hence, in practice,

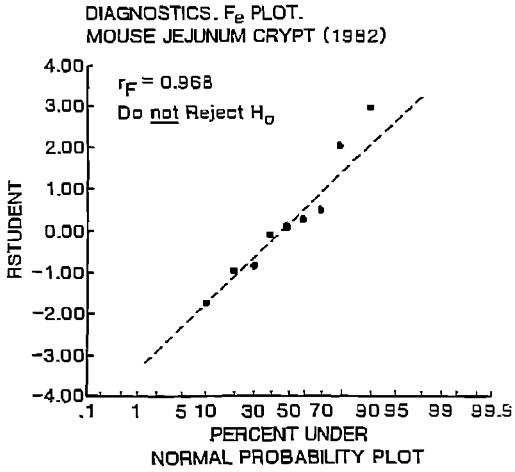


Fig. 23a. Normal probability plot of the Studentized residuals (Belsley, Kuh and Welsch, 1982) of the regression model $D^{-1} = \beta_1 + \beta_2 D/N$ of the data of the F_e-plot of Fig. 22 (where now D = D(S)). Although on the evidence of the linearity of the plot and the value of Filliben's probability plot correlation coefficient, r_F, these data do not reject the hypothesis of Normality, it is apparent that the two outlying observations may be "contaminants".

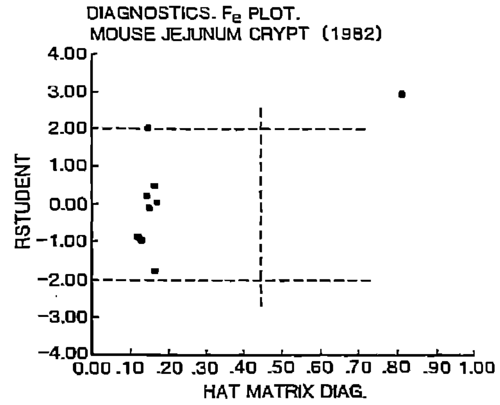


Fig. 23b. Plot of the Studentized residuals, e_i^*, vs the hat matrix diagonals, h_i, for the model $D^{-1} = \alpha + \beta D/N$ of the F_e-plot of Fig. 22. The dotted lines mark the respective "cut-offs": ± 2.00 for e_i^* and $2k/n = 0.44$ for h_i. It is evident that the observation at D/N = 15 Gy ($h_i = 0.85$) will drive the parameter estimates of the model $D^{-1} = \alpha + \beta D/N$ and functions thereof, e.g., α/β. But the stipulated limit of validity of the LQ model is $0 \leq D/N \leq 10$ Gy.

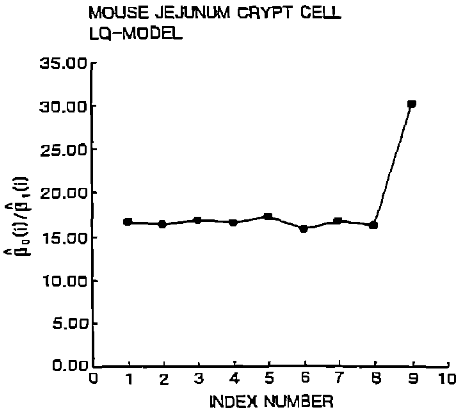


Fig. 23c. Plot of the ratio of the row-deleted estimates $\bar{\beta}_1(i)/\bar{\beta}_2(i) = \bar{\alpha}(i)/\bar{\beta}(i)$ for the linear regression model of the F_e-plot of Fig. 22. Here $\bar{\beta}_j(i)$ is the estimate of β_j, $j = 1,2$, obtained from the sample for which the ith row has been deleted. Note that the extreme observation at $D_9 = 15$ Gy - well outside the putative range of validity of the LQ model - clearly dominates the sample estimates of α/β. Since $\bar{\beta}_1(i)/\bar{\beta}_2(i) \simeq \theta = \beta_1/\beta_2 = 16.81$ for $1 \leq i \leq 8$.

Since $\theta = f(\underline{\beta}) = \beta_1/\beta_2$ is a non-linear function of the parameter $\underline{\beta}$, it is biased, $|E(\hat{\theta}) - \theta| > 0$. The size of the bias is strongly dependent on the variance of the denominator, $Var(\hat{\beta}_2)$. The so-called delta estimates of the bias and variance of θ are given by the respective Taylor series:

a) $E(\bar{\theta}) - \theta \simeq \beta_1 Var(\bar{\beta}_2)/\beta_2^3 - Cov(\bar{\beta}_1, \bar{\beta}_2)/\beta_2^2$

b) $Var(\bar{\theta}) \simeq Var(\bar{\beta}_1)/\beta_2^2 + \beta_1^2 Var(\bar{\beta}_2)/\beta_2^4 - 2\beta_1 Cov(\bar{\beta}_1, \hat{\beta}_2)/\beta_2^3$.

The bias is reduced in the so-called weighted-jackknife estimates $\bar{\theta}_i$ (Hinkley, 1977). The weighted jackknife point and interval (0.95 CL) estimates of θ for the data of Fig. 22 are $\bar{\theta} = 15.77$, (9.12, 22.42).

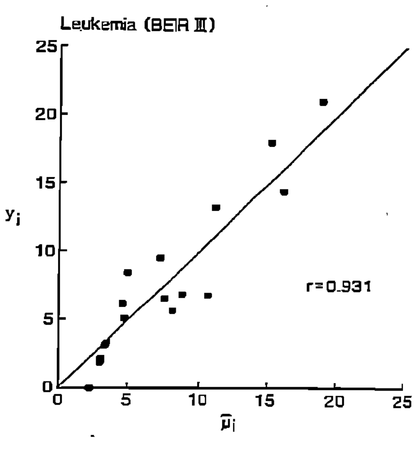


Fig. 24a. The figure presents a plot of the observed, y_i, vs expected, $\hat{\mu}_i$, responses for the LQ-L model of the BEIR III leukemia incidence data. The line of perfect fit is superimposed. "With any satisfactory model we would expect points close to and randomly scattered about this line. Any systematic deviation from the line indicates a lack of fit of the model used and gives some guide as to its nature." (Gilchrist, 1984). The LQ-L model appears to fit on this criterion. However, the chi-residuals plot of Fig. 21a for the L-L model suggests that, in fact, the LQ-L model overfits these data.

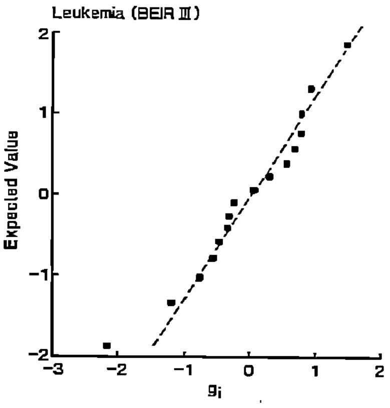


Fig. 24b. The figure presents a Normal probability plot of the Freeman-Tukey residuals, g_i, for the LQ-L model of the BEIR III leukemia incidence data. There is a single residual, g_i = -2.16 at y_i = 0, for which both its size and its location with respect to the best straight-line defined by the remaining residuals suggests that this residual identifies an "outlier". The Freeman-Tukey residuals, g_i, are more appropriate that the Pearson chi-residuals, χ_i, when Poisson data include observations at which the response is zero.

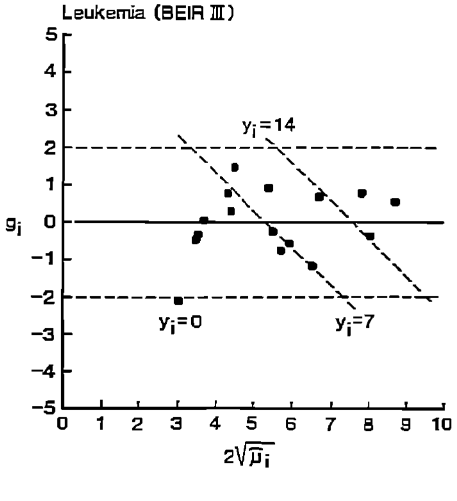


Fig. 24c. Plot of the Freeman-Tukey residuals, g_i, for the LQ-L model of the BEIR III leukemia incidence data. The residual g_i = -2.16 at y_i = 0 is clearly an outlier. The parallel lines defined by sets of similar values of y are characteristic of all residual vs fitted value plots: "Parallel lines of this nature occur no matter what model is fitted to y, and correspondingly no matter how $\hat{y}$ [=$\hat{\mu}$] is calculated, be it based on linear or non-linear estimation. So long as y-$\hat{y}$ is plotted against $\hat{y}$, the parallel lines will exist either explicitly for repeated y values, or implicitly for single y values." (Searle, 1988).

For models of Poisson responses the parallel lines have slope of -1 when g_i is plotted against $2\sqrt{\hat{\mu}_i}$, the so-called constant-information scale. (McCullagh and Nelder, 1989).

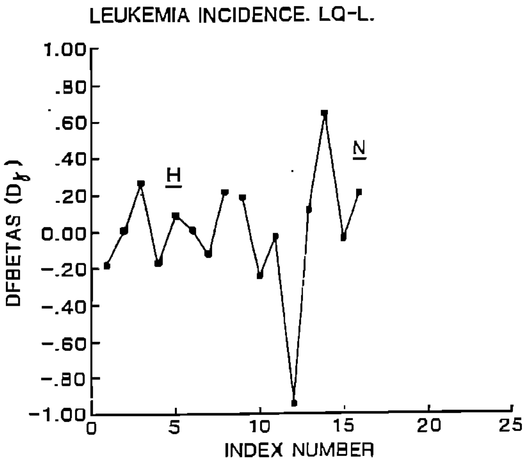


Fig. 24d. Index plot of the regression diagnostic DFBETAS for the Maximum Likelihood (ML) parameter estimate $\bar{\beta}_1$ of the LQ-L model of leukemia incidence (BEIR III). DFBETAS is a standardized measure of the respective changes in each of the components of ($\bar{\beta}$ - $\bar{\beta}$(i)), where $\bar{\beta}$(i) is the estimate of β , the parameter vector of the model, that is obtained from the sample with this ith observation (row) deleted, $1 \le i \le n$. The size-adjusted cut-off for DFBETAS is $|2/\sqrt{n}| = 0.50$. Thus, it is evident from the plot that observations #12 and #14 - especially the former - profoundly influence the ML estimate β_1; in fact, it can be shown that observation #12 provides most of the empirical evidence for choosing the LQ-L model over the L-L model of these data.

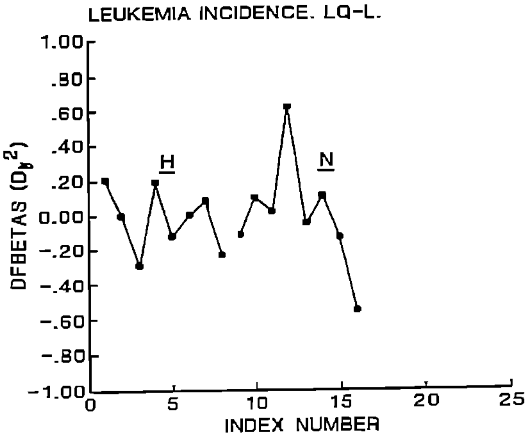


Fig. 24e. Index plot of the regression diagnostic DFBETAS for the ML parameter estimate β_2 of the LQ-L model of leukemia incidence (BEIR III). It is evident from the plot that observations #12 and #16 profoundly affect the ML estimate β_2.

Figures 23 and 24 both present evidence of samples in which the estimate of the parameter vector of the respective LQ model is dominated by a single observation. In both instances the dominant observation is somewhat anomalous: In the data of Fig. 23 the observation lies well beyond the putative range of validity of the LQ model. In the data of Fig. 24 the dominant observation - #12 - consists of a zero response at a dose well above zero: $(D_\gamma, D_n) = (38.1, 0.1)$ in rads.

It should be noted that one criterion for an adequate model of a set of data is that $\bar{\beta}(i) \simeq \bar{\beta}$, $1 \le i \le n$. Obviously, the ML estimate of β for the LQ-L model of the leukemia incidence data is quite labile and hence so are those estimates of functions of β such as $\theta = f(\beta) = \beta_1/\beta 2$, the so-called cross-over dose. The extreme lability of $\bar{\beta}$ for the LQ-L model, disclosed by the large excursions in $\bar{\beta}$ - $\bar{\beta}$(i) presented in Figs. 24a and 24b, is due to 1) the LQ-L model overfits the data, thereby inflating Var($\bar{\beta}$) and 2) the strong correlation of the terms in D_γ and D_γ^2 which also inflates Var(β). (Recall that $\bar{\beta}_{(i)} = \bar{\beta} - (X^TX)^{-1}e_ix_i^T/(1-h_i)$ for the Normal theory model.)

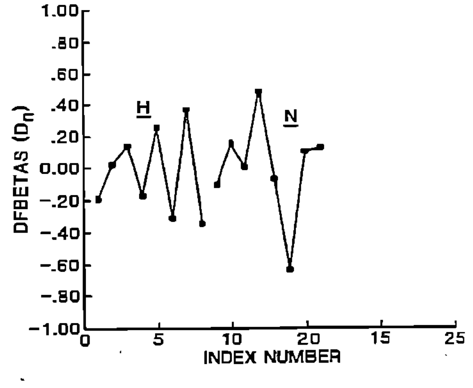


Fig. 24f. Index plot of the regression diagnostic DFBETA for ML parameter estimate $\hat{\beta}_3$ of the LQ-L model of leukemia incidence (BEIR III). The estimate, $\bar{\beta}_3$, appears to be a bit less labile than do the estimates $\bar{\beta}_1$ and $\bar{\beta}_2$. This may be attributed to the fact that the correlation between D_n and D_γ is less than that between D_γ^2 and D_γ. The pooling of the Hiroshima and Nagasaki data has reduced the correlation between D_γ and D_n - this was the object of the "data augmentation" maneuver (described in Fig. 37) - but not that between D_γ and D_γ^2. Note that the correlation between the linear and quadratic terms increases as the range of D_γ decreases.

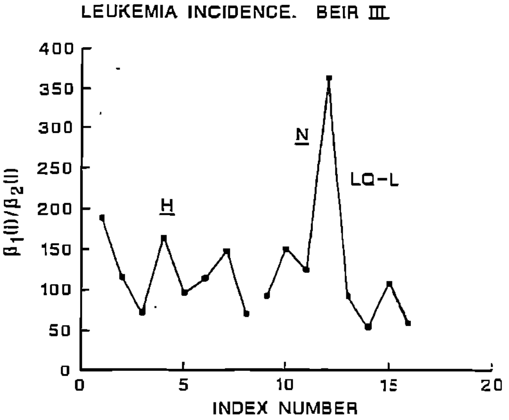


Fig. 25a. Index plot of the row-deleted estimate, $\bar{\theta}_{(i)} = \hat{\beta}_1(i)/\bar{\beta}_2(i)$, $1 \leq i \leq n$, of the cross-over dose obtained from the LQ-L model of the BEIR III leukemia incidence data. The observation at i=12 clearly dominates the sample estimate, $\bar{\theta} = \bar{\beta}_1/\hat{\beta}_2 = 117$ rad, since its deletion gives $\bar{\theta}_{(12)} = 362$ rad. Note that this effect is consistent with other evidence showing that this observation provides most of the empirical evidence for choosing the LQ-L model over the L-L model. As we have noted, this is a poor choice, indeed.

The sample estimate $\hat{\theta}$ is quite labile because of the presence of high multicollinearity in the pooled data. The bias in the estimate $\hat{\theta}$ is very large because the LQ-L model is overfit and hence Var($\bar{\beta}$) is inflated.

The weighted jackknife point estimate of θ is $\bar{\theta}_J = 17$ rad. Var($\bar{\beta}_2$) is too large to give meaningful interval estimates of θ; for the LQ model of these data, $\hat{\beta}_2/\sqrt{Var(\hat{\beta}_2)} = 1.460$.

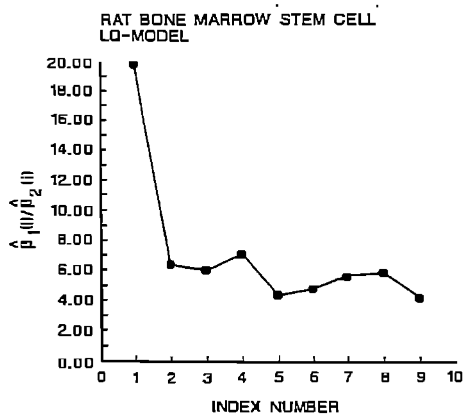


Fig. 25b. Index plot of the row-deleted estimate, $\bar{\theta}(i) = \bar{\beta}_1(i)/\hat{\beta}_2(i)$, of the α/β ratio for the LQ model of bone marrow stem cell survival data in the rat; $4 < \bar{\theta}_{(i)} < 21$, a factor of five. Obviously, the observation at i=1 (D=0) dominates the ML estimate $\hat{\theta} = 5.882$ and $\bar{\theta}_{(i)} \doteq \hat{\theta}$, $2 \leq i \leq n = 9$. The weighted-jackknife estimate is $\hat{\theta}_J = 3.246$. The weighted jackknife estimates, $\bar{\theta}_J \pm t_{0.975}(8)\sqrt{Var(\hat{\theta}_J)}$, of the 0.95 CL on $\bar{\theta}_J$ include zero. The estimate of the 0.95 CL obtained by Fieller's theorem is (1.760, 19.989). See Annex II, part 5. For the LQ model of these data, $\bar{\beta}_2/\sqrt{Var(\hat{\beta}_2)} = -4.317$.

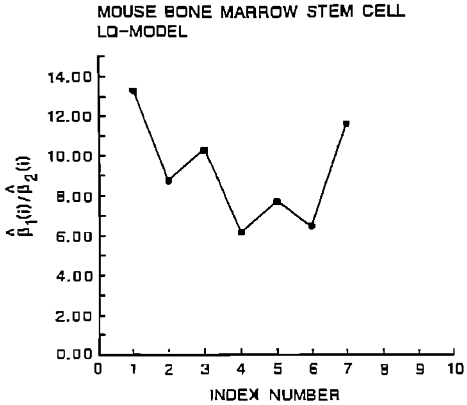


Fig. 25c. Index plot of the row-deleted estimate, $\bar{\theta}_{(i)} = \bar{\beta}_1(i)/\bar{\beta}_2(i)$, of the α/β ratio for the LQ model of bone marrow stem cell survival in the mouse; $6 < \bar{\theta}_{(i)} < 14$, a factor of two and a half. The ML estimate, $\bar{\theta} = \hat{\beta}_1/\hat{\beta}_2 = 8.902$, is obviously not dominated by any single observation as was the case for the LQ model of rat data of Fig. 5a: $\bar{\theta}_{(i)} \neq \bar{\theta}$, $1 \leq i \leq 7$. The weighted jackknife estimate is $\bar{\theta}_J = 7.619$. The weighted jackknife estimates, $\bar{\theta}_J \pm t_{0.9785}(6)\sqrt{Var(\hat{\theta}_J)}$, of the 0.95 CL on $\hat{\theta}_J$ include zero. The estimate of the 0.95 CL obtained by Fieller's theorem is (4.137, 21.538). See Annex II, part 5. For the LQ model of these data $\bar{\beta}_2/\sqrt{Var(\hat{\beta}_2)} = -4.317$.

a modeler is very apt to be mislead by the appearance of such plots, especially for small sample sizes, if a decision on goodness-of-fit is based solely on examination of the residuals (incidentally, the residuals are not independent, as there are only (n-k) degrees of freedom among them)." Nonetheless, analysis of the respective LQ models of two sets of cell survival data that is presented in Annex II, part 3, from which Figs. 19 and 20 (above) were taken, shows very well the decisive role of residual analysis. For the set of data on the rat bone marrow stem cell survival, the LQ model is narrowly rejected on the evidence of the aggregate goodness-of-fit statistic: $p = 0.04$. However, examination of the plot of χ_i residuals vs dose discloses, quite decisively, that the model has underfit the data; the presence of a pronounced 3rd order pattern in the residuals plot vividly demonstrates that the model parameter β has not captured all of the information in the data on dose response - again, some has obviously "leaked" into the residuals, χ_i. See Fig. 20.

Figures 26a and 26b are plots of the chi-residuals, χ_i, for the rival Target theory models of the mouse and rat bone marrow stem cell survival data. (Compare with Figs. 20a and 20b, respectively.) These are non-linear Poisson regression models. It is evident that for these models the evidence of the case statistics, χ_i, $1 \leq i \leq n$, should be (and was found to be) consistent with that of the respective aggregate statistics, $\Sigma\chi_i^2$. The Target theory models "fit" the data. (Although there is a weak "pattern" present in Fig. 26b, it is evident that $\chi_i < 2.0$, $1 \leq i \leq n$.)

7.10 The issue of qualitative lack-of-fit of a model.

"No single theory ever agrees with all the known facts in its domain ... It will be convenient ... to distinguish between two different kinds of agreement between theory and fact: numerical disagreement and qualitative failure."

P. Feyerabend, 1980

The results of the secondary analyses of the studies listed in Table 1 in which both the numbers at risk at each dose level and the number and distribution of dose levels were adequate (or, in the case of the distribution could be made so by a simple log transformation of the dose) offer serious challenges to some of the "facts" of the received wisdom in matters of radiation dose-response. The particular studies referred to here are 1) the radiation lethality studies (rat and mouse bone marrow stem cell survival) of Annex II, part 5; and 2) the radiation tumorigenesis studies (mammary neoplasia in female Sprague-Dawley rats) of Annex III, part 6.

It is important to distinguish here between two distinct ways in which models of dose-response may fail to "fit" the data from which they are constructed: numerical disagreements and qualitative failures. These may be referred to as metrical and topological failures, respectively. (See Zellner, 1984). "Numerical disagreement" is defined and measured by plots of the residuals and by the sum of squared residuals, for a given model of a set of data. That is, by systematic comparisons of observed and expected responses. "Qualitative", or topological failures have to do with the issues of the shape of dose-response curve of a model or, more precisely, with the sign and size of the local curvature, k, of the curve in different regions of the curve, or at different levels of dose. These are the issues of the presence or absence of "thresholds" in the dose-response curves for tumorigenesis and of "shoulders" in the dose-response curves for cell-survival that correspond to the respective models of these radiation effects. We consider these two issues separately below.

We first take up in section 7.10.1 the question of the dose-dependence of the curvature of cell survival curves. The local curvature, k, of a plane curve, $y = f(x)$ is defined as (See Courant and John, 1965.)

$$k = (d^2y/dx^2)/[1 + (dy/dx)^2]^{3/2}$$

We shall be especially interested in whether or not the dose-response curve for a given model has a low-dose "shoulder region", which may be defined in terms of two conditions on k:
i) $k < 0$ for $0 \leq D \leq D_0$, $k = 0$ at $D = D_0$, ii) $k > 0$ for $D > D_0$.

Following a discussion of the question of "shoulders" in cell-survival curves, we examine the question of "linearity" in the dose response curves for tumorigenesis in section 7.10.2. Since both of these issues are discussed at length in Annex II part 5 and Annex III part 6, respectively, we confine our remarks to brief summaries.

7.10.1 Does the cell survival curve of the single fraction LQ model have a "shoulder?"

> "A 'curvy' dose-response curve will obviously have a small value of α/β and a rather straight one will have a large value of α/β. ... The α/β ratio is useful for characterizing the response to fractionated radiation of various tissues."
>
> J. Fowler, 1984

In the first study the received model is the single fraction LQ model: $m = \exp(\beta_0 + \beta_1 D + \beta_2 D^2)$. The rival model is the Target theory model: $m = \beta_0[1-(1-e^{\beta_1 D})^{\beta_2}]$. Note that both models have the same number of parameters. They are non-nested rivals and hence the difference in the respective deviances (or Pearson chi-squares) cannot be used to discriminate between them. However, the Target theory (T) model is the better "fit" to both the rat and mouse cell-survival data on the evidence of both the aggregate and case statistics. Compare Figs. 20a vs 26a and 20b vs 26b. See also Annex II, part 5. Plots of the chi-squared residuals of the respective LQ models of both the rat and mouse data disclose the presence of strong patterns whereas, the cognate residuals plots for the respective Target theory models do not. This suggests that the LQ model has not captured all of the information on dose-response that was in these data - some has obviously "leaked" into the residuals. In the case of the LQ model of the mouse bone marrow stem cell survival, there is a positive correlation in the residuals that is suggestive of a fourth-order pattern, but the pattern is not as pronounced as is the third-order pattern in the residuals of the LQ model of the rat data.

It is also the case that we found that the parameter estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$ of the Target theory model are invariant between the rat and mouse data whereas, the cognate parameter estimates of the LQ model are not. See Annex II, part 5. And in the low-dose region, $0 \leq D \leq 2.0$, the dose-response curve of the Target theory model of the data for each species is concave-down, the curvature, k, is negative, $k < 0$, and there is, at slightly higher values of dose, a turning point, $k=0$. Thus, the Target theory (T) model describes a"shoulder region" in the dose-response curve. But for the LQ model of the same data, the dose-response curve is concave-up, $k > 0$, over the entire range of dose; therefore, there is no "shoulder region". See Figs. 27a and 27b and Annex II, part 5. But it is only the former that is consistent with received opinion on the proper "shape" - concave-down - of cell-survival curves for low LET radiation. That opinion stipulates that cellular architecture or kinetics is such that either redundant critical structures ("targets"), or innate capacities for repair of certain kinds and levels of radiation damage, are present in single cells. Both hypotheses require that the curve is concave-down in the region $0 \leq D_i \leq 2.0$ Gy. Note also that on the criterion of parsimony, the Target theory model is the model of choice, since for the same number of free parameters - 3 - it describes a more "complicated" dose-response curve, biphasic, than does the received LQ model for which the corresponding curve is monophasic.

According to an abundant literature on the matter, the LQ hypothesis invests the ratio α/β of the LQ model with both substantive meaning and practical utility: The ratio can be used to discriminate a) between early and late radiation responses, and b) between the less "curvy" from the more "curvy" cell survival curves. We now examine both propositions with some care in Fig. 27c. In Fig. 27c the curves of survival, 1S and 2S, and their respective curvatures, $1k_s$ and $2k_s$, for two LQ models of cell survival data are presented. The curves of curvature, $1k_m$ and $2k_m$, for the respective rate constants, $m_i = S_i e^{\beta_0}$, are also included. For these two sets of data the sample estimates of model parameters are $(\hat{\beta}_1, \hat{\beta}_2) = (0.469, 0.053)$ and $(\hat{\beta}_1, \hat{\beta}_2) = (0.080, 0.010)$, respectively. These models have nearly equal values of the so-called α/β ratio: $0.469/0.053 = 8.9$; $0.080/0.010 = 8.0$. However, the respective predicted survivals are obviously vastly different: The respective levels of survival at $d = 2.0$ Gy are 0.32 and 0.82. (Fertil and Malaise, 1981, 1985 have proposed using cell-survival at 2 Gy to predict tumor radiosensitivity.) Clearly, the α/β ratio cannot discriminate between these radiation responses with much finesse. Moreover, it does not appear that

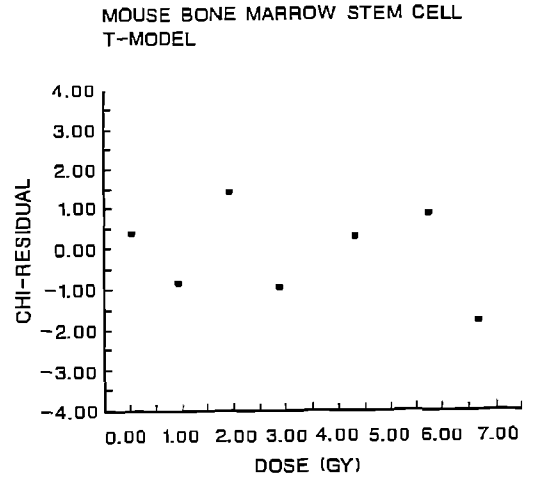


Fig. 26a. Plot of the scattergram of chi-residuals, χ_i, vs dose, D_i, for the Target theory model, $m_i = \beta_0[1-(1-e^{\beta_1 D_i})^{\beta_2}]$ of the mouse bone marrow stem cell survival data. Note the absence of serial correlation, and hence of any "pattern", in the plot suggesting that the model has captured all of the information on dose-response that is present in the data. Moreover, none of the $\chi_i > 2.0$. This suggests that the rival Target model provides a rather better fit to these data than does the received LQ model. This is consistent with the evidence of the comparison of the respective chi-squared statistics $\chi^2 = \Sigma\chi_i^2$: 7.597 vs 8.138. For both df = 4.

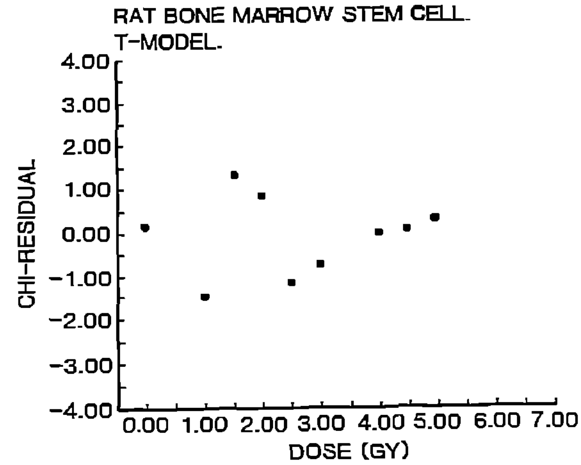


Fig. 26b. Plot of the scattergram of chi-residuals, χ_i, vs dose, D_i, for the Target theory model, $m_i = \beta_0[1-(1-e^{\beta_1 D_i})^{\beta_2}]$ of the rat bone marrow stem cell survival data. Note the absence of serial correlation and hence of pattern in the distribution. Moreover, none of the $\chi_i > 2.0$. This suggests that the rival Target model provides a better fit to these data than does the received LQ model. This is consistent with the evidence of the comparison of the respective aggregate (chi-squared) statistics $\chi^2 = \Sigma\chi_i^2$: 6.714 vs 13.214. For both models the number of degrees of freedom, df = 6.

Nelder (1968) and McCullagh and Nelder (1983) have cited invariance of the parameter vector of a model, that is, parameters that do not change as external conditions change as a criterion of its validity. And Box, Hunter and Hunter, 1978 have remarked that, "... in an adequate model, constants stay constant when variables are varied." Therefore, it is important to note that it can be shown that (for the two data sets examined) the parameters β_1 and β_2 of the Target model are relatively invariant between the two species, rat and mouse, whereas the parameters β_1 and β_2 of the cognate LQ models are not. Thus, on the criteria of concordance and invariance the Target model is the model of choice - for these data.

N.B.: The validity of between-species extrapolations of any dose-response information cannot be assumed without checking - even for two closely-related species. For example, "Of 392 chemicals in our data base tested in both rats and mice, 226 were carcinogens in at least one test, but 96 of those were positive in the mouse and negative in the rat or vice versa. This discordance occurs despite the fact that rats and mice are very closely related ..." (Ames, 1987).

α/β can readily distinguish the more from the less "curvy" survival curve either, since the curvature of the survival curve for $\alpha/\beta = 8.9$ (curve $1k_s$) is obviously much greater than that for $\alpha/\beta = 8.0$. But two basic ontological tenets of the LQ hypothesis are that the α/β ratio can be used to a) discriminate between (early and late) radiation responses and b) discriminate the more from the less "curvy" survival curves, as remarked above.

If the dose-response curves for the T and LQ models of the rat bone marrow survival data are re-plotted on the semi-log plots that are commonly used to display cell survival data, then both of the curves are concave-down, $k < 0$, in the low dose region. This strongly suggests that the "shoulder region" in dose-response curves of the LQ model of cell-survival in some data is simply a methodological artifact - a methods-instigated hypothesis. Compare Figs. 27b, 28a, and 28b.

Boag et al (1963) have described another instance of a "methods-instigated hypothesis" in radiation biology that is quite similar to this and is illustrated in Fig. 28c. In Fig. 28c the spurious curvature in the dose-response curve that is induced by the logarithmic transformation of the dose can be interpreted as empirical evidence for a "threshold" in the dose-response relationship just as the spurious curvature induced by the logarithmic transformation of the response in Fig. 28a can be interpreted as evidence for a "shoulder" in the response. See Annex II, part 5, for more complete discussion.

In the case of the LQ model of the rat bone marrow stem cell survival data the residuals plot of Fig. 20b shows that there is a strong third-order pattern in the chi-residuals of the LQ model, suggesting that a cubic term in dose, D^3, should be included. The plots of Figs. 19 and 20 suggest that, rather than a mechanistic model, the LQ model may in some instances simply be an under-fit Taylor series approximation to the true (Target theory?) model. We now examine this proposition more closely.

Figure 29 presents a super-position of the dose-response curves for the Target theory model, $m_i = \beta_0\{1-[1-\exp(\beta_1 D_i)]^{\beta_2}\}$ and the LC model, $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i^3)$ of the rat stem cell survival data. The LC model is clearly a Taylor-series approximation to the Target theory model over the range of the observations only. This implies that the LQ model is simply a Taylor series approximation to the Target theory model also. The LC model is the better approximation over the range of the data; the LQ model is a better approximation beyond the range of the observations (See Fig. 27b). But it is the former range which is of the greater clinical significance.

In Annex II, part 5, it is shown that the linear predictor, η, of the best-fitting third-order model has the form,

$$\eta = \exp(\beta_0 + \beta_2 D^2 + \beta_3 D^3)$$

It will be noted at once that for this model the slope of the dose-response curve is zero at zero dose. But one of the received prior arguments for the LQ model is, of course, that it provides a dose-response curve with negative slope at zero dose (See for instance, Alper, 1981). This suggests that in some instances the received wisdom on the boundary conditions on dose-response models could also profit from some re-thinking. We examine another such instance below.

7.10.2 Was (is) the dose-response curve for mammary neoplasia in rats "linear?"

"When the percent of rats with mammary neoplasia is plotted against the dose ... and a straight line calculated by the means of a least squares fit ... two things are apparent ... As the dose is doubled the response doubles approximately. Note that the regression lines are calculated excluding the zero dose and only up to ... 250R. Even so, the line calculated extrapolates closely to the experimentally determined response at zero dose ... It is considered that the current data confirm the previous report of Bond et al ... and the conclusions in that report that, 'the data, when analyzed in this manner, suggest that under these circumstances a threshold, if it exists, is small'."

Shellabarger et al, 1969

"Is the lower end of the dose-response curve linear or curvilinear? ... it seems very likely that the nature of the dose-response curve ... will be found to be similar across species lines."

K. Clifton, 1978

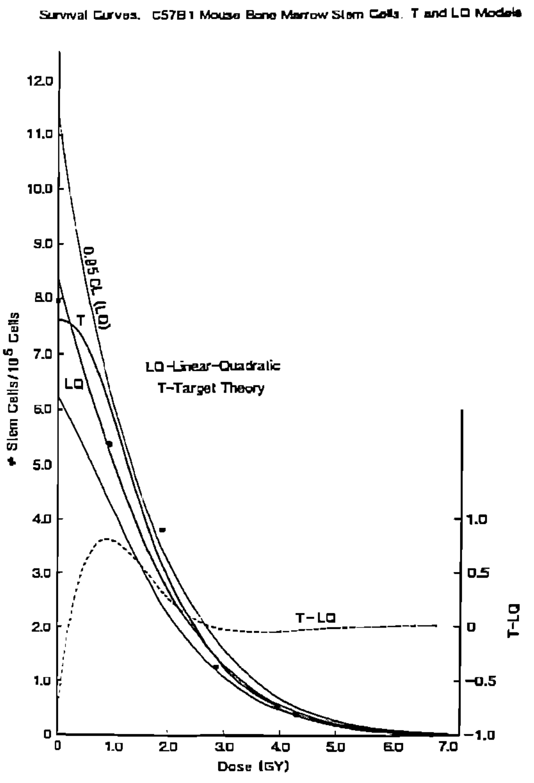


Fig. 27a. Comparison of Target theory (T) and LQ models of survival data (in vivo) of mouse bone marrow cells. Note that the greatest difference in the respective predictions of response lies in the "shoulder region". This suggests that, on the basis of the Kullbeck-Leibler divergence, or information, criterion, an experimental design to discriminate between the Target theory and LQ models will place one or more additional observations in this region (Bard, 1974), a recommendation that is intuitively correct. However, the present set of data can be used to discriminate between these two models on the two criteria of concordance (with the data) and consistency (with apriori information). (N.B.: "In any case, LQ is simply a low-dose approximation to equations that do become straight exponentials at higher doses" J. Fowler, 1989.)

The usual goodness-of-fit test based on the chi-squared approximation to the sampling distribution of an aggregate statistic, the sum of squared residuals - either χ_i or d_i - is equivocal. Both models fit about equally well on this criterion. However, comparison of the respective distributions of the case statistics, in the plots of chi-squared residuals, presented in Figs. 20a (LQ) and 26a (T) shows that the Target theory model is more consistent with the data than the LQ model. In particular, the T model fits the data better in the so-called shoulder region where the respective chi-residuals are $\chi_3 = 1.39$(T) and $\chi_3 = 2.06$(LQ). This comparison of aggregate and case statistics well illustrates the remarks of Robins and Greenland (1986) anent goodness-of-fit measures: "Another problem with goodness of fit tests is that they are insensitive to certain type of inconsistencies between models and data, and so they can indicate a good fit for certain models that are grossly inconsistent with the data. Thus, if one wishes to use a model consistent with the data, one ought to check for model adequacy by examining residuals, screening for outliers, and employing similar 'diagnostic techniques'."

It should also be remarked that the LQ model is wholly inconsistent with the apriori requirement on the curvature, k, of the survival curve in the low dose region (say $0 \leq D \leq 1.5$ Gy) - the curvature must be negative ($k < 0$) - whereas the T model is consistent with this criterion. These comparisons of the LQ and Target models well illustrates another comment of Robins and Greenland (1986), namely the recommendation to "... choose a model that is consistent with the data and yields parameter estimates consistent with ... prior beliefs."

With larger samples, say 12-15 levels of dose, the technique of embedding could be used to discriminate between these two (LQ and Target theory) non-nested rival models of cell survival. (See Gilchrist Statistical Modelling, and the discussion under Fig. 6c above.)

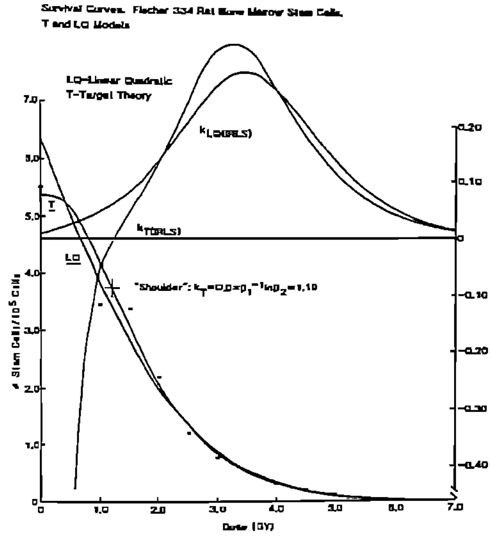


Fig. 27b. Curves of surviving number, m, and curvature, k, for the rival models, LQ and T, of the rat bone marrow data. For any curve, $y = f(x)$, the curvature $k = (d^2y/dx^2)/[1 + (dy/dx)^2]^{3/2}$. For $k < 0$ the curve is concave down. For $k > 0$ the curve is concave up. $k = 0$ at points of inflection. (A curve for which $k = 0$ at every point is a straight line.) Note that for the T model, the curve of surviving number, m, has a point of inflection, $k_T = 0$, at $D^* = \beta_1^{-1}\ln\beta_2 > 0$ and therefore has a true "shoulder". However, for the cognate curve for the LQ model, $k_{LQ} > 0$ for all $D > 0$. Therefore, the LQ model does not "see" a true "shoulder" in these data. It is for this reason that the chi-residuals for the LQ models are larger in the "shoulder region" than are the residuals for the rival Target theory models. See Figs. 20a and 20b. Of course, the appearance of a "shoulder" can be readily conferred on the LQ model simply by plotting the predicted levels of S on the familiar semi-log grid. However, it should be clear that such a shoulder is an artifact of the methodology (a "methods-instigated hypothesis"?) rather than the empirical evidence of an inherent feature of the response of the irradiated system that is captured by the Target model.

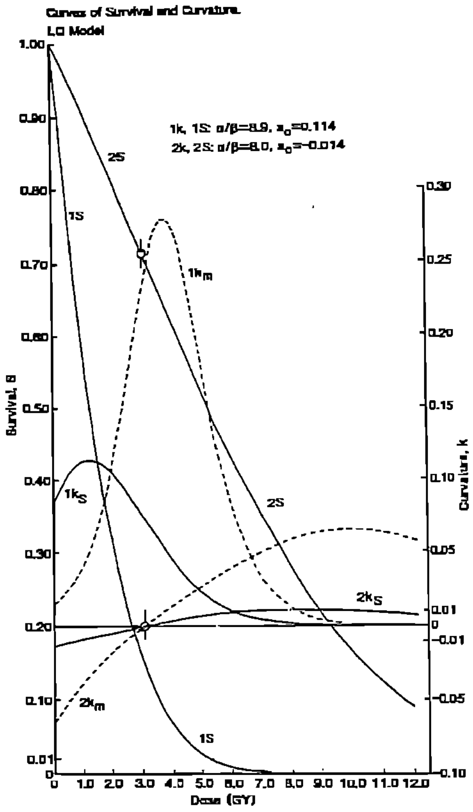


Fig. 27c. Comparison of the dose-response curves of the LQ models for two cell populations for which the respective sample estimates of α/β ($=\beta_1/\beta_2$) are similar: population 1, $\alpha/\beta = 8.9$, population 2, $\alpha/\beta = 8.0$. 1S and 2S are the respective survival curves of the two models. 1km and 2km are the curves of curvature, k, for the respective equations $m = \exp(\beta_0 + \beta_2 D + \beta_2 D^2)$ of the two models and $1k_s$ and $2k_s$ are the curves of curvature, k, for the transformed response: $m \longrightarrow me^{\beta_0} = S$. (The curvature of a plane curve $y = f(x)$ is defined as $k = (d^2y/dx^2)/[1 + (dy/dx)^2]^{3/2}$. See Fig. 27b.)

It is evident from the figure that the respective survival curves of two cell populations with quite similar values of α/β may differ markedly. For the example of Fig. 27c the respective ratios, α/β, only differ by about 10%, while the respective surviving fractions at, say, 2Gy, differ by about a factor of 2.0. Discrepancies as great as this should not be unexpected since the ratio, α/β, is, of course, unique only to within a multiplicative factor. However, the fact that a ratio, α/β, should have been chosen to discriminate between early ($\alpha/\beta = 3.0$) and late ($\alpha/\beta = 11.0$) radiation responses in normal tissue is most unexpected.

A different function, a_0, of the parameter vector of the LQ model, based on the curvature, k, of the associated dose-response curve has some value as a criterion for discriminating between survival curves. It is discussed in Annex II, part 5 of the present report. The criterion is qualitative rather than quantitative and discriminates between these curves for which a shoulder is present ($a_0 < 0$) and those for which it is not ($a_0 > 0$).

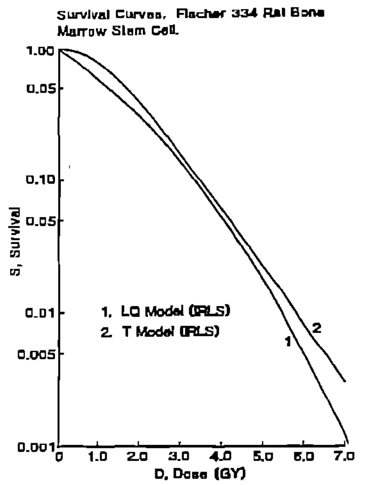


Fig. 28a. Semi-log plots of the survival curves, S, of the respective dose-response models, T and LQ, of the rat data. Note that both curves are concave down - curvature k < 0 - in the low dose region suggesting the presence of a "shoulder" in the radiation response. Comparison with the cognate dose-response curves of Fig. 27b discloses that this feature is a methodological (or representational) artifact for the LQ model. For the LQ model the logarithmic transformation of the response, S ⟶ log S, has reversed the sign of the curvature of the survival curve: k > 0 ⟶ k < 0. However, for the T model the sign of the curvature is "invariant under the log transformation" suggesting that it is an inherent feature of the process described by that model.

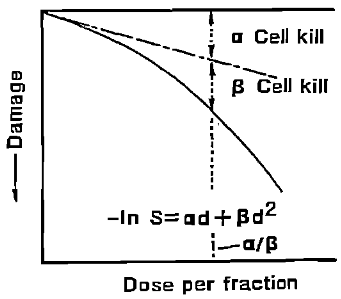


Fig. 28b. "The linear quadratic description of dose-response curves. The ratio α/β (grays) is the dose at which the effects of the linear and quadratic terms become equal. α/β is small if the curve is 'curvy' and large if it is relatively straight." (Reproduced with permission from Fowler, 1984). Replacing the adjective "curvy" by the more precise description provided by "curvature", $k = (d^2y/dx^2)/[1 + (dy/dx)^2]^{3/2}$, shows that the ratio α/β can be a misleading criterion for the "shape" of a survival curve. Examination of Fig. 27c discloses that the curvature, k, for the survival curve with α/β = 8.9 is greater - and hence the curve is "curvier" - than that for the survival curve with α/β = 8.0. Compare curve $1k_S$ with curve $2k_S$. Note that the survival curve 2S has both positive and negative curvatures but that both are quite small.

Thus, we see that the α/β ratio is a rather ambiguous criterion for the shape of a cell-survival curve as well as for the discrimination between the early and late effects in tissues.

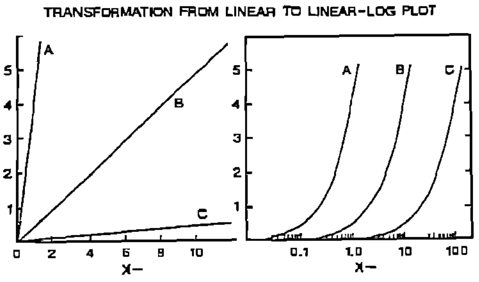


Fig. 28c. Comparison of arithmetric vs semi-log plots of (hypothetical) dose-incidence curves (Reproduced with permission from Boag et al, 1959). Incidence is described by the ordinate and dose by the abscissa in each plot. Note that the logarithmic transformation of dose has induced a positive curvature in the dose-incidence curves that can be - and has been - interpreted as empirical evidence for the presence of a biological feature - a "threshold dose". This is, of course, completely analogous to the negative curvature induced in the survival curve of Fig. 28a by the logarithmic transformation of the response that is presented in Fig. 27b which has been interpreted as empirical evidence for the presence of another biological feature, a "shoulder". But in each case, the biological feature is an artifact of the respective logarithmic transformation: D ⟶ logD and m ⟶ logm. "The greatest shortcoming of the human race is man's inability to understand the exponential function" (A. Bartlett, 1976). Or its transform.

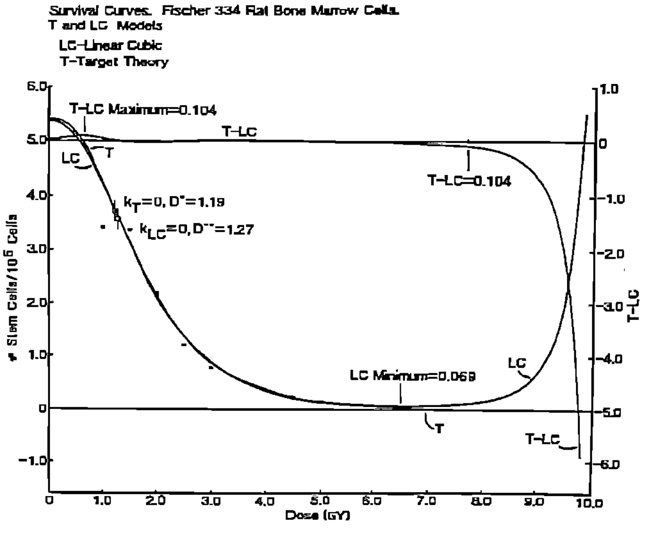


Fig. 29. Comparison of the estimates of the Poisson rate constant m_i by the rival LC and T models of the rat cell survival data: LC. $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i^3)$, T, $m_i = \beta_0[1-\{1-\exp(\beta_1 D_i\}^{\beta_2}]$. The plot of the chi-residuals, χi, of the LQ model of this data in Fig. 20b displayed a marked cubic pattern. Addition of a term in D_i^3 to give the model LC improved the fit significantly as measured by the decrement χ^2 between respective chi-squared statistics, $\chi^2 = \Sigma\chi_i^2$. The LC model obviously interpolates well, especially in the so-called shoulder region for which the survival curve is now "concave-down", consistent with prior information, whereas the survival curve for the cognate LQ model is "concave-up" in this region (See Fig. 27). However, it is evident that the LC model is completely unsuitable for extrapolation beyond the region of the data. The figure well illustrates the weakness of the Osiander criterion and the advantages of a mechanistic over an empirical model.

"Concerning x-ray and gamma-ray radiation, some investigators have concluded that the dose-response relationship appears to be linear (Bond et al, 1960b; Shellabarger et al, 1969; ... Moskalev and Petrovieh, 1972; Hellman et al, 1982), although rigorous statistical analyses have not been used to exclude other forms of dose-response relationships."

Shellabarger et al, 1986

We now compare the non-nested rival linear and probit models of radiation-induced mammary neoplasia in the female Sprague-Dawley rat that is examined in Annex III, part 6. The received linear model of tumor incidence rate is $\pi = \alpha_0 + \alpha_1 D$, where $0 \leq \pi \leq 1.0$, is the observed proportion of responders at dose D (Shellabarger et al, 1969). The original data, together with the linear dose-response curve published by Shellabarger et al in 1969, is presented in Fig. 13a. Figure 30a presents the UNSCEAR (1977) graph of these data together with a curve "connecting" - not "fitting" - the observations.

The linear model does not provide an adequate estimate of the spontaneous incidence rate, C, (which is known a priori to be non-zero) since the 0.95 confidence limits on α_0 include 0. The rival probit model is $z = \beta_0 + \beta_1 x_1$; C, where the probit transform is $z = \Phi^{-1}(\pi)$, $\Phi(\ )$ is the Normal distribution function, $x_1 = \log D$ and C is the spontaneous incidence rate (at D=0); the 0.95 confidence limits on C do not include 0. (N.B.: We wrote a computer program that provided maximum likelihood estimates of β_0, β_1, and C following the methods described in Finney, 1971b.) The dose-response curve for this model is shown in Fig. 30b. The probit model "fits" the better of these two rivals on the evidence of both aggregate and case statistics. Indeed, since the rival, probit model, $z = \beta_0 + \beta_1 x_1$; C, does provide both point and interval estimates of C, the received linear, non-threshold model may be said to underfit these data. It is also the case that the transformation of dose, D   logD, as well as the transformation of response, π z, that is required by the probit model, achieves a more useful representation of the data of the Shellabarger et al experiment since it reduces the leverage of the observation at D = 250R. This is shown in Fig. 30c which presents a super-position of the respective index plots of Cook's distance, D_i, for the linear, $\pi = \alpha_0 + \alpha_1 D$, and probit, $z = \beta_0 + \beta_1 \log D$; C, models of the Shellabarger et al data. Moreover, the probit model has the greater "scope", since, in order that the linear, non-threshold, model "fit" the X-ray data, the responses at D=0 and D=500 rad must be omitted. Such omissions are not necessary to fit the probit model. See Fig. 31 (and Annex III, part 6).

Figure 31a is a super-position of the respective dose-response curves for the linear and probit models of the Shellabarger et al data. The dose-response curve is, of course, the distribution function of the tolerance dose that corresponds to a given model. Figures 31b and 31c present plots of the respective density and distribution functions for the tolerance doses implied by the linear and probit models of the Shellabarger et al data. Clearly, the rectangular distribution of tolerance dose that is implied by the linear model is inconsistent with the a priori information on the issue, whereas the Normal distribution of (log) tolerance dose that is implied by the probit model is not.

The secondary analysis of the Montour et al 1977 data of Fig. 13b that is described in Annex III, part 6, disclosed that the estimate, $\hat{\beta}_1$ of the slope parameter of the probit model is invariant between sets of dose-response data in which the LET of the radiation is different - neutron vs γ-ray. That is, $\hat{\beta}_1(\gamma) = \bar{\beta}(n)$. This invariance of slope is shown quite vividly in Fig. 32. The corresponding sample estimate of neutron RBE is given by 10^{θ} where $\hat{\theta} = \bar{\beta}_2/\hat{\beta}_1$ from the probit model $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ of the pooled γ-ray and neutron data described above (and in Annex III, part 6). Note that the invariance of β_1 with LET implies that the neutron RBE, for this response or end-point, is independent of dose. This is, of course, contrary to the received wisdom on this matter, which holds that the neutron RBE varies inversely with the level of dose. Since the rival probit models "fit" these data better than do the received linear models, this suggests that the issue of the dose-dependence of neutron RBE should be re-examined.

Shellabarger et al remark that "... the regression lines are calculated excluding the zero dose and only up to 200 or 250R. Even so, the line calculated extrapolates closely to the experimentally determined response at zero dose." But as the line A in Fig. 33a discloses, it extrapolates still closer to zero response at zero level. One recalls the treatment of neutron RBE for the L-L model of breast cancer in the BEIR III (1980) report in which it was reported that the confidence limits on the coefficient on D_n included the estimate of the coefficient of D_γ - but not that these confidence limits also included zero. (See Herbert, 1986c)

It should be noted that the BEIR III Q-L model of breast cancer, for which the dose-response curve on D_γ is, like the curve on D_γ of the probit model of mammary neoplasia, "concave-up" in the low-dose region, does in fact "fit" the LSS (T65D) data: For the Q-L model $P(\chi^2 > 17.03|16) = 0.384$. For the L-L model $P(\chi^2 > 8.40|16) = 0.936$, where 17.03 and 8.40 are, respectively, the chi-squared goodness-of-fit statistics for the Q-L and L-L models. But, for the latter, the goodness-of-fit statistic lies in the lower 0.06 tail - the degree of "fit" is almost too good to have arisen by random sampling from a population described by the model. (See section 7.3.2. and Fisher, 1963.) This provides a rather different interpretation to the BEIR III statement that, "... the best-fitting function linear in both D_γ and D_n corresponded to a chi-square statistic for lack of fit that was half as large as that obtained from the best-fitting model linear in D_γ^2 and D_N." Moreover, the ratios $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)} > 1.90$ for both D_γ and D_n, which is consistent with the T65D dosimetry - unlike the L-L model. This tends to refute the BEIR III statement that, "Breast cancer data offer little support for a dose-response model with a strong upward curvature in D_γ." Finally, we must remark on still another statement in the BEIR III (1980) report in support of the committee's choice of the L-L model of breast cancer induction: "Linear model coefficients for D_γ and D_N did not differ significantly." This is true, of course; however, it has been shown (Herbert 1986c) that the coefficient for D_N in the linear model also does not differ significantly from zero.

On the other hand, it is shown in Annex III, part 6, for the probit model of the Shellabarger et al data, $z = \beta_0 + \beta_1 x_1$; C, that the 0.95 CL on $\hat{C}$, the estimate of the spontaneous incidence do not include zero but that the 0.95 CL on the intercept, $\hat{\alpha}_0$, of the linear model, $\pi = \alpha_0 + \alpha_1 D$, do (see also Fig. 30b).

From Fig. 33 it is evident that the dose-response curve for the probit model of the mammary neoplasia data is bi-phasic. In the low-dose region the slope of the curve approaches zero. (Thus, the similarity of the curves in Figs. 30a and 33b is quite striking and yet the former was quite consistently interpreted as being consistent with a linear, non-threshold model of dose-response!) However, the dose-response curve for the (received) linear model has a constant slope. On the criterion of parsimony the probit model is (like the Target model of cell-survival) the more parsimonious since for the same number of free parameters (2) it describes a more "complicated" dose-response curve - biphasic - than does the received linear model for which the corresponding curve is monophasic. Compare Figs. 30b and 31a with Fig. 13a, b (See also Annex III, part 6). Moreover, the shape of the dose-response curve for the probit model is evidently more consistent with a large body of other empirical evidence than is that of the received (linear) model: "Ionizing radiation can transform normal cells to a latent tumour stage but it can also facilitate the growth of such transformed cells, particularly at higher doses. The latter effect is probably a result of radiation damage to the anti-tumour defense It is thus not surprising that the experimental dose-tumour relationship often indicated a bi-phasic course with a low-dose limb, the slope of which cannot be distinguished from zero and a high-dose limb that increases with the dose ... There are four or five exceptions to this 'rule' but in most of thee exceptional cases we know the probable reasons for the deviations. Unfortunately, the 'exceptional cases' have been so extensively reported and discussed that people sometimes believe that they represent the normal animal tumour response to radiation. This is not the case." (G. Walinder, 1978).

However, one of the accepted prior arguments for the received linear model is that it has a dose-response curve with a positive slope at zero dose, i.e., there is no non-zero threshold. (See also Annex III, part 6.) Note that the received models for both cell-survival and carcinogenesis stipulate similar boundary conditions on the slope of the respective dose-response curves, namely, that the slope is non-zero at zero dose.

The ontological implications that emerge from the logarithmic transformations deployed in sections 7.10.1 and 7.10.2. should be noted:

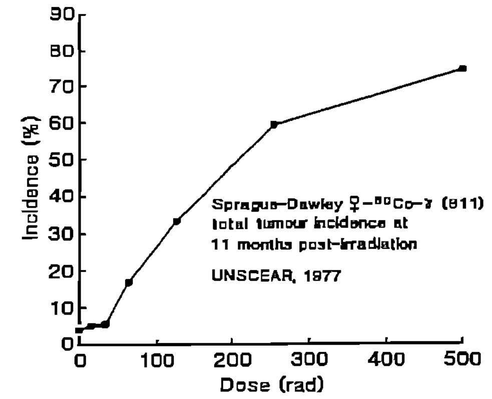


Fig. 30a. The figure presents the Shellabarger et al (1969) data of Fig. 13a above. It is reproduced (with permission) from Figure VII of the UNSCEAR 1977 report in which it is observed that, "126. The Sprague-Dawley rat is the animal where most studies on the dose-related induction of breast tumours have been carried out (fig. VII). … A fairly good linear relationship with exposure of the observed percentage incidence of tumour-bearing animals … was found in the range 16-250R for both types [x or gamma] of radiation. From this observation it was concluded … that a threshold did not exist, the data extrapolating satisfactorily to control incidence" … "127. The evidence reviewed allows the conclusion that the dose effect relationship for mammary tumour induction in the rat appears to be linear down to very low doses."

The UNSCEAR 1986 report notes that "… the pattern seems to emerge that the shape of dose-response relationships may be basically similar in various species. 458. Thus, the linear non-threshold type of dose-response relationship for female mammary carcinoma induced by x-rays finds its counter part in the results of studies on at least four strains of rats: the Sprague-Dawley females irradiated externally by x-rays and/or gamma rays, …

NCRP 64, 1980 observes that, "In these animals, total body … ^{60}CO gamma irradiation … the incidence scored within one year after exposure increases as a linear function of the exposure … from 16-200R (Shellabarger et al, 1969) …"

The BEIR III 1980 report remarks that, "Breast-cancer data offer little support for a dose-response model with strong upward curvature in D_γ. The dose-response curves for mammary tumors in female rats given total-body x and gamma irradiation tend to be linear."

Our secondary analyses of the Shellabarger et al (1969) data (see Fig. 13a) disclose quite clearly that "none of the above" statements are correct.

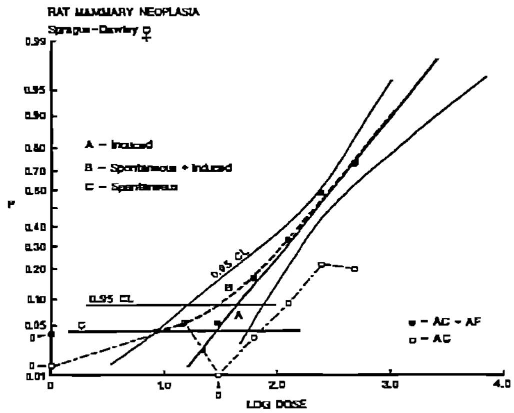


Fig. 30b. Plot of the dose-response curve for the probit models, $z_i = \beta_0 + \beta_1 x_1$; C, of the incidence of mammary neoplasia (AC + AF) and mammary cancer (AC) in female Sprague-Dawley rats exposed to Co^{60} radiation, where $z = \Phi^{-1}(\pi)$ - probit of the response, $0 \leq \pi \leq 1$, $\Phi(.)$ - Normal distribution function, x_1 - log dose, C - spontaneous incidence rate, AC - adeno carcinoma, AF - adeno fibroma. These are the Shellabarger et al, 1969 data of Figs. 13a and 30a. The 0.95 confidence limits (CL) for (AC + AF) are superimposed. Note that the rival probit model, curve B, is consistent with all 7 observations in the Shellabarger et al, 1969 data, whereas, the received linear model $\pi = \alpha_0 + \alpha_1 D$, curve A, is only consistent with those five observations in the region $0 < D \leq 250$. (As a matter of fact, Shellabarger et al, 1969, deleted the observations at $D_1 = 0$, $D_7 = 500$, before fitting the linear model; a maneuver that recalls Acton's 1959 comment: 'Some data fit lines naturally, but others have lines thrust upon them.")

The probit model suggests the presence of a non-zero threshold dose, $D_0 \doteq 25$ rad, in the (AC + AF) response. Note that the respective received models for cell survival and mammary neoplasia (LQ and linear, respectively) give dose-response curves with non-zero slopes at D=0, while the respective rival models of these two responses give dose-response curves with near-zero slopes at D=0.

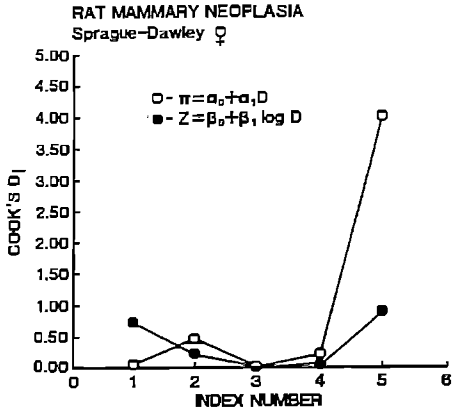


Fig. 30c. Super-position of the index plots of the Cook's D_i diagnostics for two rival models of the Shellabarger et al (1969) data on radiation-induced mammary neoplasia in Sprague-Dawley female rats in the region $0 < D \leq 250$. The linear model, $\pi = \alpha_0 + \alpha_1 D$, $0 \leq \pi \leq 1$, is that fitted by Shellabarger et al (1969). However, the probit model, $z = \beta_0 + \beta_1 x$, $z = \Phi^{-1}(\pi)$, where $\Phi(\ )$ is the standard Normal distribution function and $x = \log D$ is, of course, the correct model for a binary response. It will be noted that the doses in the Shellabarger et al (1969) experiment are distributed geometrically (see Fig. 13a) which is appropriate for a model on the log dose metameter, but for the dose metameter it assures that the sample estimate of the parameter vector will be dominated by the observation at the highest dose, as shown in the plot of Cook's D_i. Since the interest is usually in the response at the lower doses this is the source of still another weakness in the Shellabarger et al interpretation of their data. It can be seen in the plot of Cook's D_i that for the probit model on the log dose metameter no single observation dominates the sample estimates of the parameter vector.

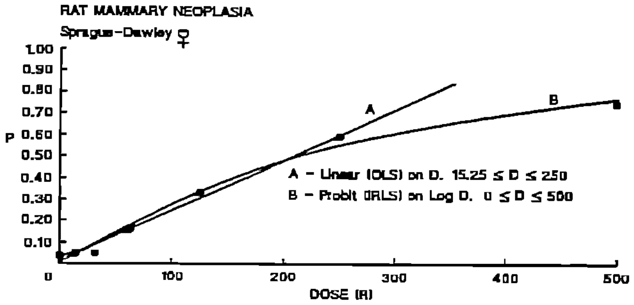


Fig. 31a. Super-position of the dose-response curves for the received linear, $\pi = \alpha_0 + \alpha_1 D$, and the rival probit, $z = \beta_0 + \beta_1 x_1$; C, models of the mammary neoplasia data (AC + AF) of Fig. 30 where $0 \leq \pi \leq 1$ is the response, $z = \Phi^{-1}(\pi)$ is the probit of the response, $\Phi(.)$ is the Normal distribution function and $x_1 = \log D$. Note that the data to which the received linear model was fit was reduced by deletion of the original observations at D=0 and D=500 rad, as in the Shellabarger et al, 1969 paper, while the rival probit model was fit to all of the original observations. The dose-response curves describe the respective cumulative distribution functions for the two tolerance distributions. For the received linear model the tolerance distribution is uniform, or rectangular; for the rival probit model the tolerance distribution is Normal (or, more precisely, log-Normal). But only the latter tolerance distribution is consistent with apriori information on the matter. See UNSCEAR 1986: "The susceptibility of an irradiated animal or human population to tumor induction is assumed to follow a bell-shaped distribution."

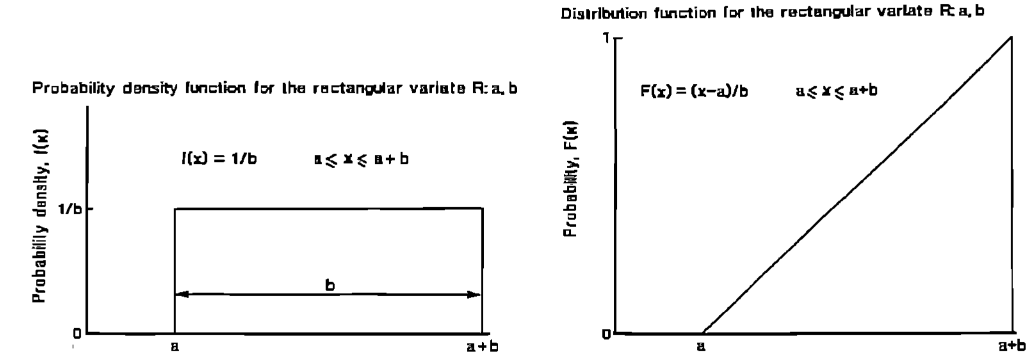


Fig. 31b. Probability density and distribution functions for the Uniform distribution. The Uniform distribution describes the tolerance distribution for the linear model, $\pi = \alpha_0 + \alpha_1 x_1$, $0 \leq \pi \leq 1$, of a binary response such as a stochastic radiation effect.

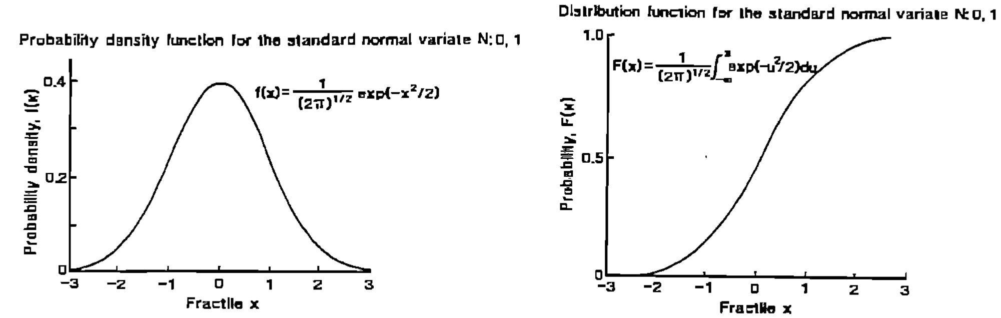


Fig. 31c. Probability density and distribution functions for the Normal distribution. The Normal distribution describes the tolerance distribution for the probit model, $z = \beta_0 + \beta_1 x_1$, $z = \Phi^{-1}(\pi)$, $0 \leq \pi \leq 1$, of a binary response such as a stochastic radiation effect, where $\Phi(.)$ is the standard Normal distribution function. The Figs. 31b and 31c are redrawn from N.A.J. Hastings and J.B. Peacock, Statistical Distributions.

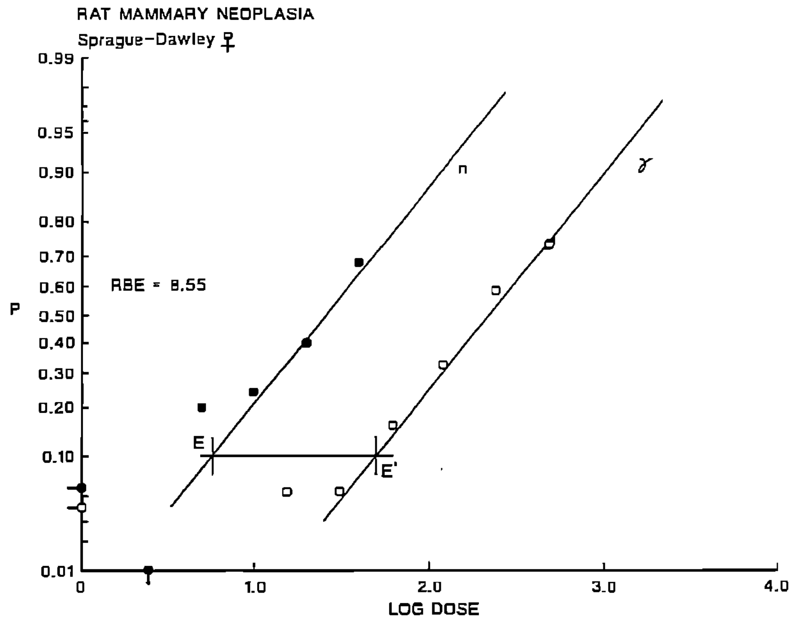


Fig. 32. Super-position of the respective dose-response curves for the probit models, $z = \beta_0 + \beta_1 x_1$; C, of the Shellabarger et al, 1969, (Co^{60} γ-radiation) and the Montour et al, 1977 (35 Mev (max) neutron radiation) data on mammary neoplasia (AC + AF) in Sprague-Dawley female rats. The respective equations are:

a) Co^{60} $z = -4.583 + 1.951x_1$; $C = 0.042$ $n = 7$
(-5.064) (5.180) (2.133)#

b) neutron $z = -2.72 + 1.960x_1$; $C = 0.045$ $n = 6$
(-3.008)(2.948) (2.175)#

The slope, β_1, is obviously invariant between these two sets of data. The spontaneous incidence rate, C, is also common to the two sets. Therefore, point and interval estimates of the neutron RBE can be obtained from the probit model $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where x_2 is the indicator variable

$$x_2 = \begin{cases} 0 - \text{neutron} \\ 1 - \text{gamma} \end{cases}$$

The neutron RBE is 10^{θ} where $\theta = \beta_2/\beta_1$. For these data RBE $\simeq$ 9 compared with the received estimate RBE $\simeq$ 4 (Montour et al, 1977).

Note that this secondary analysis discloses that, contrary to the received wisdom on the matter, the neutron RBE is independent of dose for this response. It may, of course, depend on the dose for other responses. However, Montour et al 1977 reported that for their data, shown above, the neutron RBE is dependent on dose.

$\beta_j/\sqrt{Var(\beta_j)}$

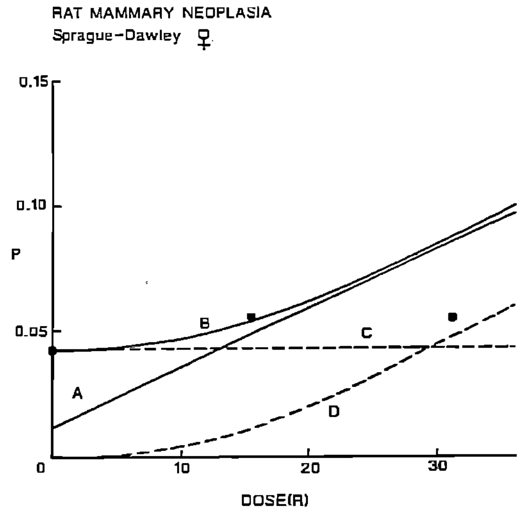


Fig. 33a. Comparison of the respective dose-response curves of the received linear, curve A, and rival probit, curves B, C, and D, models of the rat mammary neoplasia data of Shellabarger et al, 1969. Curve B is the sum of curves C and D. Note that the probit model provides a better description of the observed response at D=0 than does the linear model. For these data the observation at D=0 is quite precise. Recall from Fig. 32 that the probit model also provides a better description of the observed response at D=500 rad than does the linear model. Therefore, on the criterion of scope the probit model is the greater. ("An important property of a model is its scope, i.e., the range of conditions over which it gives good predictions. Scope is hard to formalize but easy to recognize, and intuitively it is clear that scope and parsimony are to some extent related _ Both scope and parsimony are related to parameter invariance, that is, to parameter values that either do not change as some external condition changes or change in a predictable way" (McCullagh and Nelder, 1983).)

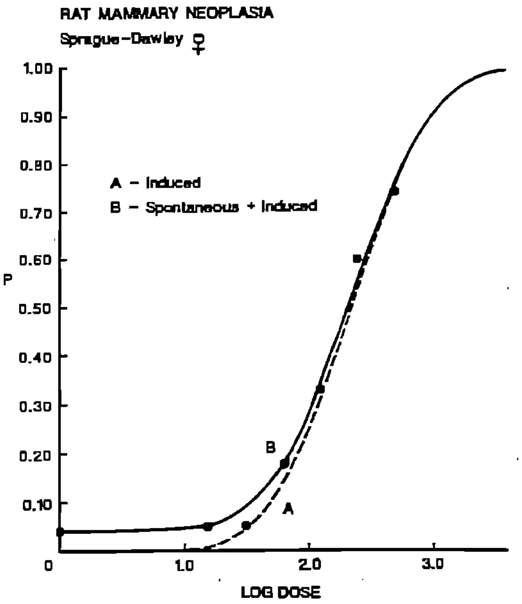


Fig. 33b. Super-position of the plots of the dose-response curves of the probit models of spontaneous + induced incidence vs induced incidence of the rat mammary neoplasia data of Shellabarger et al, 1969 on a semi-log plot. It is important to remark that the sample estimate of the parameter vector β of the probit model, $z = \beta_0 + \beta_1 \log D$, of the six observations in the Shellabarger et al data for which $D > 0$ can be shown to be (nearly) invariant with respect to deletion of the observation at D = 500R, i.e., $\bar{\beta}_{(6)} = \beta$. This suggests that these data are homogeneous in the sense that the observed levels of response describe a single process - tumour induction - rather than the difference of two competing processes - tumour induction + cell-killing - as has been often suggested in the literature. ("In an adequate model, constants stay constant, when variables are varied." Box, Hunter, and Hunter, 1978)

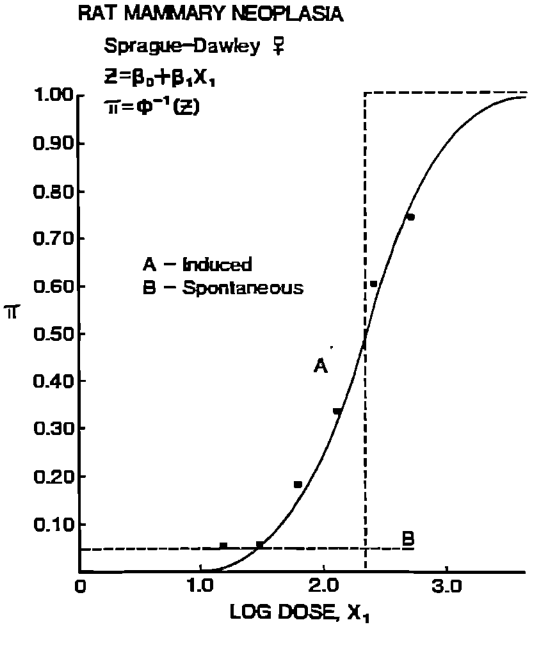


Fig. 33c. The graph is a super-position of the plot of the curve A of Fig. 33b which is the distribution function for the Normal distribution of tolerance and the distribution function of the δ-function distribution of tolerance which it approximates. The difference between the two distribution functions is due to the dispersion of "tolerance doses" in the members of the sample about the "threshold dose" defined by the step function. As is well-known. The reciprocal, say σ^{-1}, of the measure of dispersion of the tolerance dose that is due to "biological variability" is a measure of the slope of the dose-response curve the more homogeneous the population the steeper is the slope. This is inherent in all tolerance dose models which are therefore, intrinsically "threshold" models. An effective threshold or E-NOEL (effective no observed effect limit is defined by the intersection of curves C and D in Fig. 33a or by the intersection of curves A and C of Fig. 30b.

i) For the LQ model of cell-survival the transformation S $\longrightarrow$ logS of the fitted curve leads to a dose-response curve with a negative curvature, k < 0, that is currently interpreted as implying the existence of either cellular architectures or cellular processes which are not "seen" (by the model) - and hence may not exist - for this model, except as artifacts of the log transformation of the fitted curve.
ii) For the probit model of radiotumorigenesis the transformation D $\longrightarrow$ logD of the data $(r_i/n_i, D_i)$ leads to invariance of the slope parameter between low-LET and high-LET exposures as shown in Fig. 32.

We have shown that there are consequential contradictions between the sui generis surrogate methods and criteria of estimation and inference by which the results that are regularly reported in the current radiobiological literature are obtained and the normative methods of statistical modelling. This suggests that these results should be re-examined. The perdurability of the surrogate methods and criteria, despite the demonstrable inferiority of their estimates and inferences to those provided by statistically adequate methods and criteria, may be partly explained by Zipf's Principle of Least Effort (Zipf, 1965). We will discuss this principle in greater detail below. However, this perdurability (of these concepts, etc.) can perhaps also be partly accounted for by two other of Kuhn's observations anent paradigm-based scientific research: 1) "... the mature sciences are regularly more insulated from the external climate, at least of ideas, than are other creative fields" (Kuhn, 1977) and, 2) "... the members of a given scientific community provide the only audience and only judges of that community's work" (Kuhn, 1962/1970).

Thus, in the insulated praxis of radiation oncology, the construction and assessment and deployment of models of dose-response have been developed quite independently of the concepts, methods, and criteria of statistics and in particular of statistical modelling. The ideas of the latter have been marginalized, where not ignored altogether. The latter, apparently, are regarded by most investigators as ideas belonging to an altogether different creative field - a field called Statistics.[9] But that notion is, of course, quite wide of the mark. Since the field of statistics is concerned both "with the logic of scientific method and with how we learn from data" (S. Fienberg, 1989), the two fields are virtually coincident. Or, more precisely, statistical methods provide the canonical set of procedures that help the scientist to assess the strength of the evidence supplied by a sample of data for or against a hypothesis and to assess the reliability of an estimate - or recommendation - obtained from such a sample (Pitman, 1979). Indeed, "... it is possible to maintain ... that statistics in its broadest sense is the matrix of all experimental science and is consequently a branch of scientific method, if not Scientific Method itself; and hence that it transcends the application of the scientific method in sundry fields of specialization. The scientist should know statistics as he knows logic and formal language for communicating his ideas" (M. Kendall, 1968). More specifically, a few scientists have made a good argument that biological science, when it is practiced as Science, should properly be regarded merely as a branch of Applied Statistics. To paraphrase Lord Rutherford's views on Physics: "All Science is either Statistics or it's stamp-collecting!"

7.11 Methods for reducing the effects of multicollinearity.

In section 6 it was pointed out that the presence of multicollinearity in the distribution of the predictor variables, X, led to inflation of the estimates $\underline{\hat{\beta}}$ and Var($\underline{\hat{\beta}}$) for a regression model of [$\underline{y}$, X]. Moreover, one or more of the estimates, $\hat{\beta}_j$, might also have the 'wrong sign', i.e., be inconsistent with prior information on the sign, as well as on the size of β_j.

There are two measures that may be taken to reduce the effects of multicollinearity in X on estimates of $\underline{\beta}$ depending upon whether the data [$\underline{y}$, X] are to be obtained in experimental or non-experimental studies:

1) If the data are experimental then the distribution of X in the design can usually be chosen so that the predictor variables are, more or less, orthogonal; that is, chosen so that $X_j^T X_k = 0$, $j \neq k$, where X_j, $1 \leq j \leq p$, is the column vector representing the j^{th} predictor variable. In this event, the effects of multicollinearity are reduced by eliminating the cause itself. (See Myers, 1971.)

2) If the data are non-experimental, then the effects of multicollinearity in X on the sample estimate of the parameter vector $\underline{\beta}$ can be reduced by post-hoc salvage maneuvers, in which the semi-Bayesian methods of Mixed estimation and Ridge regression and the simple meta analytic procedure of Data augmentation that were discussed above in section 7.2 are deployed. See also Annex IV, parts 3 and 5. (See Theil, 1973; Montgomery and Peck, 1982.)

We consider first the pre-hoc methods of orthogonal experimental designs. We then examine the post-hoc semi-Bayesian methods. It is important to remember that both the efficient pre-hoc design of experiments and the post-hoc salvage of sample data require good a priori information on the parameter vector, $\underline{\beta}$, of the model at issue. Both requirements are especially true for non-linear models in general and for the generalized linear models of Binomial and Poisson responses in particular.

7.11.1 Pre-hoc methods and criteria. Design.

"We design an experiment by choosing in some rational way the values of $\underline{x}$ at which $\underline{y}$ is to be observed."

Y. Bard, 1974

"Carefully designed experiments are necessary. There are no fitting techniques which can overcome the deficiencies of poorly designed experiments."

J. Blakemore et al, 1963

"Our consensus is ... that the most critical statistical errors involve improper research design and a common failure to report critical information regarding the design. Whereas one can correct incorrect analytic techniques with a simple re-analysis of the data, an error in research design is almost always fatal to the study - one cannot correct for it subsequent to data collection. For example, the presence of selection bias because of the method used to choose subjects will invalidate the conclusions for the entire population, or, if the sample size is insufficient to produce adequate statistical information, no test can compensate for the shortfall."

R. G. Marks et al, 1988

"If we are to suppose that effective design is possible at all, we must also assume therefore that something is known about the values of the parameters in advance."

G.E.P. Box and H. Lucas, 1959

"An efficient design requires good advance estimates of the parameters."

W. Cochran, 1973

".... if we take data at points chosen haphazardly in the space of the variables, then the estimates may be imprecise, or we may not be able to obtain separate estimates of the individual parameters at all."

G.E.P. Box, 1960

We remarked earlier on several weaknesses of the design of the experimental study described in Tucker and Thames (1983), e.g., the small numbers at risk, n_i, and their effects on the findings of the study, e.g., the questionable validity of the chi-squared approximation to the sampling distribution of RSS when there were excessive numbers of extreme responses; the presence of observations that dominate the fit and the parameter estimates, etc. We now consider several additional weaknesses of that study and their respective effects on study findings. We first note that "... if one is interested in learning from the experience of others, it is important to determine what problem they were attempting to solve. Upon careful examination, many apparent errors appear to represent the deft resolution of the wrong problem." (Fischhoff, 1982). Figures 9 and 10 - and the statement that $n_i = 10$, $1 \leq i \leq n = 19$ - give little hint of the nature of the problem that motivated the design of the experiment of Fig. 9. (For example, the 19 levels of observed response in Fig. 9 were collapsed into the 7 levels of estimates of D(0.50) and D(0.50)/N of Fig. 10 before any calculations are made.) However, it seems to be the case that in any report in the burgeoning literature on the LQ model, either for N=1 or $N \geq 1$ or, for that matter, for any of the received models of dose-response, it is difficult to determine that the design and analysis of the study were motivated by concern to obtain a solution to any of the more insistent problems of model construction and testing. These are usually taken to be four (Bard, 1974): a) The estimation of the parameters in a given model to specified degrees of accuracy and precision. b) The estimation of the levels of response over a specified region of the treatment variables. (Or, conversely (inversely), the estimation of the levels of treatment variables, say $\underline{\hat{x}}^T(\pi)$, to educe a specified level of response, π. Both predictions, of course, depend on the estimates of the values of one or more unknown parameters.) c) The discrimination between several rival models as to which is most concordant with "reality". d) The selection of a course of action in circumstances in which the optimal action depends jointly on what the correct model is and what the values of the parameters are. But, all of the related experiments that are reported in the literature appear to be similar to that described in Fig. 9; they seem to be motivated and informed largely by the Baconian recommendation to "vex Nature", or to simply, "Twist the lion's tail" (Bacon, 1620/1960). Or, in a contemporary locution, "Sock it and see!"

Currently, the more insistent, albeit apparently largely unrecognized, problems in radiation biology, as is the case in the present analysis, appear to be those of parameter estimation and model discrimination and the design of an experiment must of course be informed and directed by the requirements of the problem to be solved. We have described elsewhere in this report a model selection criterion, the AIC, that is derived from information theory (Akaike, 1977; 1985). It is also the case that the criteria for the optimal design in each of the four problems a)-d) above can be derived from information theory: The distribution of treatment variables within each design is selected to maximize the information obtained from the experiment (Bard, 1974). For example, one criterion for design of an experiment to obtain optimal estimates of model parameters is to choose X, the design matrix, to minimize the volume of the confidence ellipsoid on $\underline{\beta}$. This is equivalent to maximizing the determinant $\det(X^TX)^{-1}$ where X is the design matrix, the so-called D-optimal design. It can be shown that a design that maximizes this determinant also minimizes the maximum variance of the estimated levels of response, say $\underline{x}_i{}^T\underline{\hat{\beta}}$, over that region of X (the so-called G-optimal design) (Montgomery and Peck, 1982).

In general, of course, the dependence of the level of observed response, say y_i, on the levels of several treatment variables such as dose, D, fractions, N, and time, T (as well as, perhaps, such covariates or, co-factors, as sex, race, etc.) is known a priori. This information dictates the construction of multivariable models that explicitly include all such variables, ab initio, rather than by such post-hoc salvage maneuvers as we have described. And in particular, such models would lead naturally to experimental designs that are quite different from that described in, say, Fig. 9. Indeed, it would appear that such failures of experimental design as are represented by Fig. 9 are one of the immediate penalties exacted by the failure to include T explicitly in the LQ model, as remarked above under ontological weaknesses in section 6.1. For the optimal design of an experiment to obtain estimates of the parameter vector $\underline{\beta}$ of a dose-response surface - thereby further elucidating the respective roles of the several factors - is, of course, not simply a collage of one-dimensional designs for the estimation of the parameters of dose-response curves such as appears in Fig. 9. For example, the optimal design must assure that the treatment variables are uncorrelated, or only weakly correlated; that is, the design must be orthogonal, design criteria that were discussed in section 7.1.

And in designing an experiment for a Binomial response, the number, m, of levels of the treatment vector, $\underline{x}^T$, their spacing, $\Delta\underline{x}^T$, and the sample size, n_i, at each level of $\underline{x}_i{}^T$ must be selected as carefully as the range and correlation parameters of the joint distribution of the treatment variables D, N, and T; Finney, 1971a, b provides a discussion and recommendations for the selection of optimal levels of m, $\Delta\underline{x}^T$, and n_i.

Figure 34a presents a scattergram of the design of the experiment of Fig. 9 in the space of the treatment variables (D, D^2/N) for the probit model of the LQ hypothesis. The basic principle of experimental design given by Bard (1974) appears to have been well and truly violated: The "values of $\underline{x}$ at which $\underline{y}$ is to be observed" do not appear to have been chosen in any "rational way".

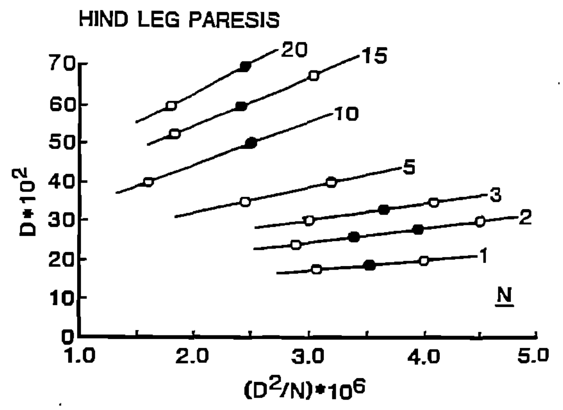


Fig. 34a. Scattergram of treatment regimens for model M2 in the plane of the variables D and D^2/N. Symbols as in Figs. 14a and 14b. The correlation coefficient is $r = -0.571$. A common criterion for experimental designs for parameter estimation is to maximize the determinant of X^TX (a D-optimal design). Obviously, the design of the experiment of Fig. 34a does not achieve this desideratum for the LQ hypothesis.

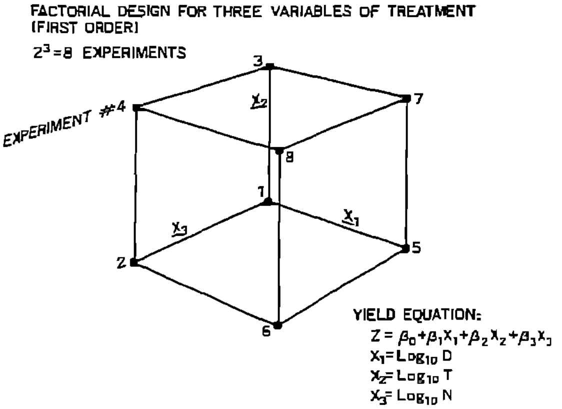


Fig. 34b The figure presents an isometric plot of a factorial experiment to estimate a response surface model for a radiation effect in tissue. The orthogonal design includes 3 factors (D,T,N) each at 2 levels. There are $2^3 = 8$ treatment regimens required to estimate the coefficients in the yield equation of first degree. Since the design is orthogonal, the separate roles of the three treatment variables can be identified.

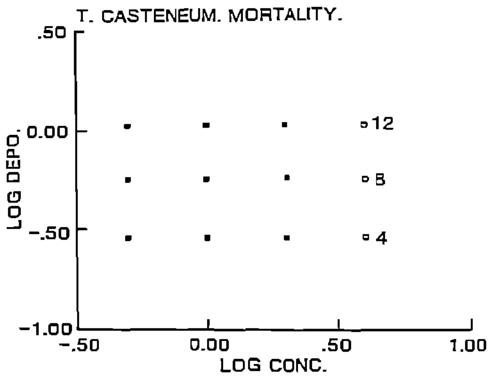


Fig. 34c. Scattergram of the distribution of treatment regimens for insect mortality experiment in log Dep. - log Conc. plane where Conc. and Dep. refer, respectively, to the concentration and total weight of the deposit of pesticide. The data are taken from Finney's Probit Analysis (1971). Note that the log-transformation has no effect on the degree of orthogonality of the treatment variables. Note also that the treatment variables Conc. and Dep. are rather analogous to the variables D/N and D, respectively, in radiation therapy.

An orthogonal design for a proposed experiment for the estimation of parameters for a model such as M1a that is linear in the logarithms of three treatment variable, here D, N and T, is presented in Fig. 34b. (See Davies, 1967; Myers, 1971.) Often more than two levels of each variable will be specified in order to detect second-order effects in the response as in the case of a design for an actual experiment that is shown in Fig. 34c. For these data the model fitted is $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $z = \Phi^{-1}(\pi)$, $0 \leq \pi \leq 1$. The data for this design are taken from Finney, 1971. The symbols in Fig. 34c have the same meaning as in Figs. 34a and 34b: there are three extreme responses, $r_i = n_i$ in the twelve treatment regimens shown in the figure. The average number at risk is $n_i = 27.6$, $1 \leq i \leq 12$. The standard deviation is $s(n_i) = 3.3$ ($\max(n_i) = 30$; $\min(n_i) = 17$). Note that in this experiment there are only $3/12 = 0.25$ extreme responses vice $12/19 = 0.63$ in the experiment of Fig. 9 and an average of $n_i = 28$ subjects at risk at each level of treatment vice 10 for the experiment in Fig. 9. (Actually, of course, there were only $n_i = 5$ at risk in the experiment that was done by van der Kogel; $n_i = 10$ was the number reported by Tucker and Thames.)

7.11.2 Post-hoc maneuvers and criteria. Salvage.

"... when the predictor variables are highly correlated, ridge regression produces coefficients which predict and extrapolate better than least squares and is a safe procedure for selecting variables."

D. Marquardt et al, 1975

"We argue that it is typically true that there is available a priori information about the parameters and that this may be exploited to give improved, and sometimes substantially improved, estimates."

D. Lindley et al, 1971

The statements of Blakemore et al (1964) anent weak experimental data that are cited in part 7.11.1 above, although still an excellent prescription, now hold with somewhat less force with arrival of several methods for post hoc salvage of certain weak studies: There now are several "fitting techniques" which can often substantially ameliorate - although not completely overcome - the effects on the sample estimates of model parameter vectors of certain of the "deficiencies" in weak non-experimental data. If the weakness arises from the presence of multicollinearity in the set of p predictor variables described by the n*p matrix X then the non-least squares techniques of Ridge regression and Mixed estimation, discussed in parts 7.2.1 and 7.2.3 above can provide substantially improved estimates of the parameter vector of the model over those provided by other fitting techniques such as Ordinary Least squares. Both Ridge regression and Mixed estimation give estimates of the parameter vector of a regression model that are linear functions of the Ordinary Least Squares estimates of the parameter vector of the model that are obtained from the sample.

In general, it may be said that Ridge regression methods are deployed when the a priori information on the parameter vector, β, as well as the cognate information in the multicollinear sample, is weak; for example, Ridge regression is often used when no more is known, a priori, than the sign and significance of the parameters, β_j, and the identity of the (correlated) predictor variables. (Ridge regression methods were developed by Hoerl and Kennard (1970) to provide better - reduced variance - estimators than those obtained by Least Squares methods from multicollinear data. See Montgomery and Peck, 1982. On the other hand, Mixed estimation can be deployed when the a priori information on β is stronger, say the specific values of either the parameters β_j, or functions of the parameters $f(\beta)$, e.g., a ratio β_j/β_k, are known a priori. Mixed estimation is also deployed to correct weaknesses in the sample data other than those due to multicollinearity. Extensive discussions of Ridge regression and Mixed estimation, illustrated with data from several of the studies listed in Table 1, are presented in Annexes II-IV. See also Herbert, 1985a, Herbert 1986a, and Herbert, 1986c.

Examples of the deployment of the techniques of Ridge regression and Mixed estimation in the construction of isoeffect models of the two sets of highly collinear, non-experimental (clinical), radiation toxicity data that are described in Figs. 4a and 4b are presented in Figs. 35 and 36, respectively. The confidence <u>ellipses</u> (0.90 CL, 0.95 CL) in Figs. 35a and 36a are constructed on the ordinary least squares estimates ($\hat{\alpha}_1$, $\hat{\alpha}_2$) of the parameter vector $\underline{\alpha}^T = (\alpha_0, \alpha_1, \alpha_2)$ of the <u>isoeffect</u> models, $x_1(\pi) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$ of the data of Figs. 4a and 4b, respectively. Here $x_1(\pi) = \log D(\pi)$, $x_2 = \log N$ and $x_3 = \log T$. It will be noted that for both sets of data the ordinary least squares estimates of $\underline{\alpha}$ "have the wrong sign" and the covariance matrix $Var(\underline{\alpha})$ has been greatly inflated by the large correlation between N and T that is present in the respective data sets.

Figure 35b is the <u>ridge trace</u>, a plot of the ridge estimate, $\underline{\hat{\alpha}}_R$, vs k, the biasing parameter, for the isoeffect model of the data of Fig. 4b. Figure 35c is a plot of the variance inflation factor, VIF vs k, for the isoeffect model of the data of Fig. 4b. Figure 35d presents a plot in the (α_1, α_2)-plane of the ridge trace of Fig. 35d. In Fig. 35c the Marquardt criterion (VIF $\sim$ 1) for choosing an "optimal" value of k, the biasing parameter, is shown. In Fig. 35d the Obenchain criterion ($\underline{\hat{\alpha}}_R$ within the 0.90 confidence ellipse on $\underline{\bar{\alpha}}$) for choosing an "optimal" value of k is compared with the Marquardt criterion. See Marquardt and Snee 1975 and Obenchain 1977, 1980.

In Fig. 36a the 7 a priori estimates of α_1 and α_2 selected by Supe et al (1983) on the basis of the related British Institute of Radiology study are super-imposed on the 0.90 and 0.95 confidence ellipses on the OLS estimates of α_1 and α_2 for the isoeffect model of the data of Fig. 4a. Note that all of these estimates lie outside the 0.90 confidence ellipse and hence can be rejected on the Obenchain criterion. Figure 36a also shows the positions of a row-deleted estimate, $\underline{\bar{\alpha}}(4)$, and the Ridge regression (RR) estimates for the isoeffect model of the data of Fig. 4a for comparison.

Figure 36b presents the super-position of the <u>ridge trace</u> for the ridge regression estimates of $\underline{\alpha}$ on the 0.50 and 0.90 confidence ellipses on the OLS estimates of $\underline{\alpha}$ for the isoeffect model of the data of Fig. 4a. The Figure also shows the positions of the Mixed estimates of $\underline{\alpha}$ for five selections of the dispersion matrix, ψ, of the a priori information represented by the constraint $\underline{r} = R\underline{\beta} + \underline{v}$, $E(\underline{v}) = \underline{0}$, $Var(\underline{v}) = \psi$. ψ specifies the level of uncertainty in the a priori information. As Robins and Greenland (1986) recommend, it is good practice to determine the <u>sensitivity</u> of the posterior estimates of $\underline{\beta}$. "... to moderate changes in ones prior beliefs ...", etc. But, as Leamer (1986) has remarked, a priori information on the parameter vector is usually easier to specify than is the precision with which that information is known. Hence, in a <u>sensitivity analysis</u>, we determined the effect on the posterior estimate of $\underline{\beta}$ of various levels of the prior ψ in the constraint $\underline{r} = R\underline{\beta}$.

It is important to note that the Supe et al 1983 study, described in Figs. 36a and 36b, which has <u>not</u> been cited even once since publication (See Table 1), is no weaker than other studies such as Thames et al 1982, Tucker and Thames 1983, Fowler 1984, Fertil and Malaise 1981, which have been cited far more often. (These latter studies have, of course, been anatomized and discussed in several of the preceding sections and in Annexes II-IV of this report.)

In addition to Ridge regression and Mixed estimation another post-hoc salvage technique for reducing the effects of multicollinearity on parameter estimates is that of Data augmentation described in part <u>7.2.2</u>. Extensive discussions of Data augmentation, illustrated with data from several of the studies listed in Table <u>1</u> are presented in Annexes II-IV.

An example of the fruitful deployment of Data augmentation techniques in the construction of <u>dose-response</u> models of highly collinear, non-experimental radiation carcinogenicity data, such as described by the within-city data (Hiroshima and Nagasaki) from the LSS sample of the BEIR III report (1980) is illustrated in Fig. 37. This technique was successful in this case because the respective correlation structures of the city-specific data were <u>complementary</u>. (It is of interest to note that the presence of the high degree of multicollinearity in the city-specific data has <u>never</u> been acknowledged as a reason for <u>pooling</u> the LSS data, although unless these data are pooled sensible parameter estimates cannot be obtained for any of the three rival models considered in the BEIR III report: L-L, Q-L, or LQ-L. Note also that such pooling provides an elementary example of weak studies "borrowing strength" from one another - as in a meta-analysis. NAS/NRC, 1992.)

An example of a dubious deployment of the Mixed estimation technique in which sample

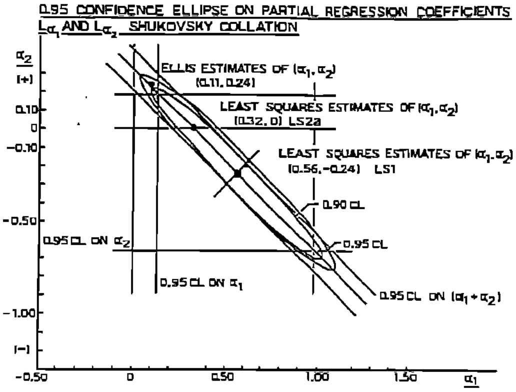


Fig. 35a. The figure presents the marginal (rectangle) and joint (ellipse) confidence regions on $\underline{\alpha}$ in the (α_1, α_2) plane where $\underline{\alpha}$ is the parameter vector of the isoeffect model of the data on tumor ablation in cancer of the oropharynx presented in Fig. 4b: $x_1(P) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$, where $x_1(P) = \log D(P)$, $x_2 = \log T$, $x_3 = \log N$. The 0.95 confidence limits on the sum ($\alpha_1 + \alpha_2$) are also shown as is the 0.90 confidence ellipse on (α_1, α_2). The ordinary least squares estimates ($\hat{\alpha}_1$, $\hat{\alpha}_2$) of (α_1, α_2) are shown as LS1. The constrained least squares estimates (constraint) $\alpha_2 = 0$) are shown as LS2a. The apriori estimates are those of Ellis (1969).

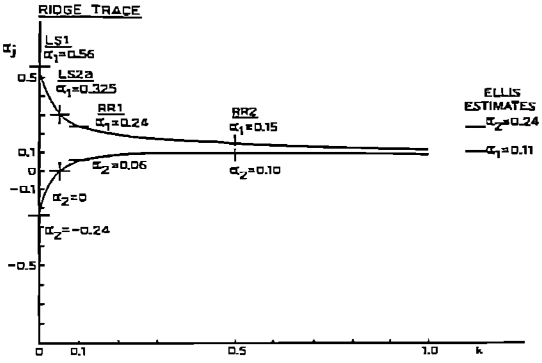


Fig. 35b. The figure presents the plot of the ridge estimate, $\hat{\underline{\alpha}}_R$ - the ridge trace - function of the biasing parameter k. The estimates, $\hat{\underline{\alpha}}_R$, for k = 0.1 (Marquardt 1970 criterion) and k = 0.5 (Obenchain 1977 criterion) are shown as RR1 and RR2, respectively. The ridge estimate can be written as a linear transformation of the least squares estimate:

$$\hat{\underline{\alpha}}_R = [X_0^T X_0 + kI]^{-1} X_0^T X_0$$

where X_0 is the centered and standardized design matrix; i.e., $X_0^T X_0$ is the correlation matrix of the treatment variables.

Ridge regression is a post-hoc salvage maneuver for collinear data that is an alternative to data augmentation (See Fig. 37).

The least-squares estimate $\hat{\underline{\alpha}}$ is, of course, $\hat{\underline{\alpha}} = (X_0^T X_0)^{-1} X_0^T \underline{y}_0$ where $X_0^T X_0$, $X_0^T \underline{y}_0$ are correlation matrices. (Obviously, if multicollinearity present in X_0 then $(X_0 X_0)$ is nearly singular and the estimates $\hat{\underline{\alpha}}$ and Var($\hat{\underline{\alpha}}$) will be inflated.)

Note that the Ellis estimates, obtained by introspection, differ appreciably from the Ridge regression estimates: 0.15 vs 0.11 or 0.24 vs 0.11 and 0.06 vs 0.24 or 0.10 vs 0.24.

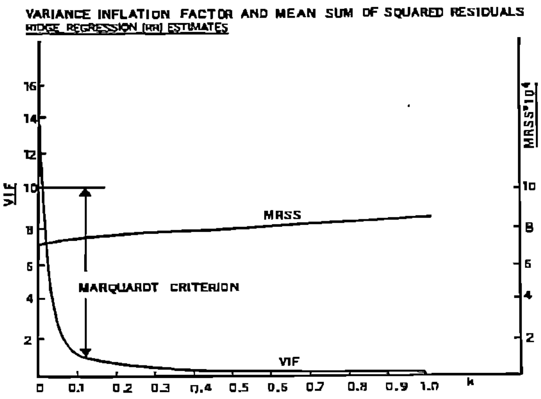


Fig. 35c. The Fig. presents plots of the mean residual sum of squares (MRSS = RSS/(n-2)) and the variance inflation factor (VIF) as a function of the biasing parameter k of the Ridge regression technique. The Marquardt criterion for optimal k is defined as VIF(k) = 1.0.

The residual sum of squares is given as $RSS^{**} = RSS + (\hat{\underline{\alpha}}_R - \hat{\underline{\alpha}})^T X^T X (\hat{\underline{\alpha}}_R - \hat{\underline{\alpha}})$ where $RSS = (\underline{y} - X\hat{\underline{\alpha}})^T(\underline{y} - X\hat{\underline{\alpha}})$ and $\hat{\underline{\alpha}} = (X^T X)^{-1} X^T \underline{y}$. The variance inflation factor is the leading diagonal element of $Var(\hat{\underline{\alpha}}_R) = (X_0^T X_0 + kI)^{-1} X_0^T X_0 (X_0^T X_0 + kI)^{-1}$ where $C = X_0^T X_0$, the correlation matrix. The variance inflation factor is obviously a much more sensitive function of k than is the mean residual sum of squares. This is one reason why ridge regression is a useful technique.

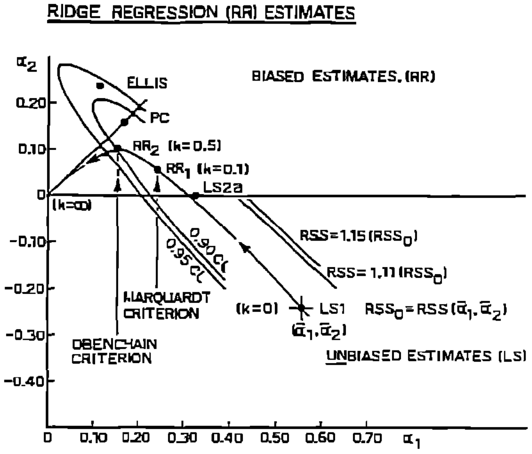


Fig. 35d. The figure presents the ridge trace, $(\hat{\underline{\alpha}}_R)$, vs the biasing parameter k, in the (α_1, α_2) space for the Ridge regression estimates of the parameter vector, $\underline{\alpha}$, of the isoeffect model of tumor ablation in cancer of the oropharynx: $x_1(\pi) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$ where x_1 - logD, x_2 - logN, x_3 - logT, and $\pi = 0.70$. As k varies between 0 and ∞, the (posterior) Ridge estimate $\underline{\beta}_R$ varies between $\underline{\alpha}$ and 0. Two of the several criteria for the selection of "optimal" values of the biasing parameter: a) The Marquardt criterion is that VIF(k) = 1.0 and gives k = 0.1 (See Fig. 35c). b) The Obenchain estimate of optimal k (=0.5) is obtained as the intersection of the ridge trace and the 0.90 confidence ellipse on $\underline{\alpha}$. Note that, as remarked in Fig. 35b, the Ellis estimates are markedly different from the Ridge regression estimates.

The ridge trace is defined by two extremum conditions: "As k increases without bound from zero, $[\hat{\underline{\alpha}}_R]$ traces a curved path from $[\underline{\alpha}]$ to $\underline{0}$. The path is uniquely determined so that the distance from $[\underline{\alpha}]$ to $\underline{0}$ is diminished as rapidly as possible while the residual sum of squares increases as slowly as possible." (Swindel, 1976). Thus, the trace moves parallel to the major and then parallel to the minor axes of the family of confidence ellipses.

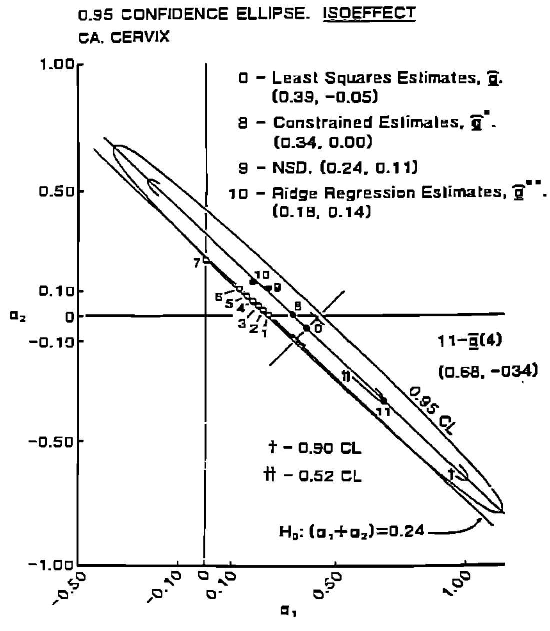


Fig. 36a. The figure presents the confidence ellipse in the (α_1, α_2)-plane on the sample estimate of the parameter vector α for the isoeffect model, $x_1(\pi) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$, of the data on the radiation treatment of cervical cancer in Fig. 4a. Here x_1 = logD, x_2 = logN, x_3 = logT, $0 \leq \pi \leq 1$. For those data π = 0.55, i.e., the model describes the 55% isoeffect curve for tumor ablation. The 0.90 and 0.95 confidence limits on the sample estimate ($\bar{\alpha}_1$, $\bar{\alpha}_2$) are superimposed.

The points 1-7 describe the several a priori estimates of obtained from the British Institute of Radiology study on cancer of the laryngo-pharynx that were deployed by Supe et al, 1983 in their published report. These all satisfy the prior constraint: $(\alpha_1 + \alpha_2)$ = 0.24. Point 10 represents the NSD estimate of α_1 and α_2 for connective tissue "tolerance". Point 11 describes $\bar{\alpha}(4)$, the row-deleted estimate of α, i.e., that estimate obtained when the observation with index number i=4 was omitted from the sample. For this observation Cook's D_4 = 1.0, and, as expected, the estimate $\alpha(4)$ is shifted to the $\simeq$ 0.50 confidence ellipse.

Note that because of the high degree of collinearity in N and T, shown in Fig. 4a, the sample estimate of α is very labile; $(\bar{\alpha} - \bar{\alpha}_{(i)}) \neq 0$, $1 \leq i \leq 8$. Note also that the apriori estimates of α all lie on or beyond the 0.95 confidence ellipse. But on the Obenchain criterion (1977), any estimates of α that lie beyond the 0.90 ellipse on the sample estimate, $\hat{\alpha}$, should be of little interest to the investigator. It can also be shown (see Annex IV, part 3) that none of these seven apriori estimates are compatible with the sample estimate on the criterion of the γ-statistic of Mixed estimation (which is distributed as Pearson's χ^2 on 2 df).

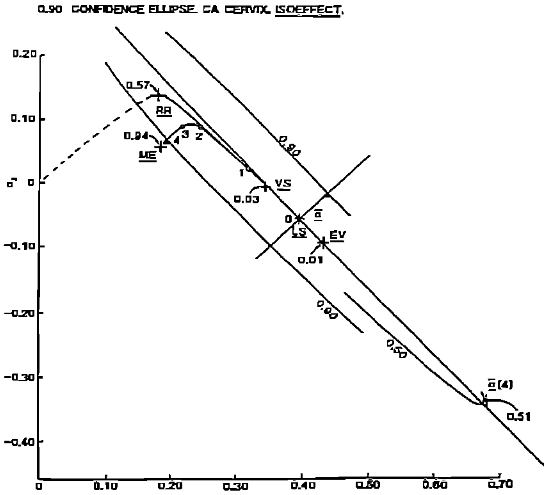


Fig. 36b. The Fig. presents the ridge trace of the Ridge regression estimates of α for the isoeffect model of the data of Fig. 4a. This is the dashed line terminating at $(\alpha_1, \alpha_2) = (0, 0)$ for $k \longrightarrow \infty$. The Ridge regression estimate (RR) is shown for k = 0.125: $\bar{\alpha}_R^T$ = (0.18, 0.14). This estimate lies on the 0.57 confidence ellipse on α.

The Fig. also presents the Mixed estimates of α obtained with the prior estimate of (α_1, α_2) = (0.18, 0.06) that was obtained by Supe et al (1983) from the British Institute of Radiology (BIR) study (1978, 1981) on cancer of the laryngo-pharynx and described by point 4 of Fig. 36a. The points at 1, 2, 3, 4, ME described the Mixed estimates, α for the following values of the precision matrix, ψ, of the prior information, $r = R\alpha + v$, $E(v) = 0$, $Var(v) = \psi$:

$$\psi = 10^{-1}I_2,\ 10^{-2}I_2,\ 10^{-3}I_2,\ 10^{-4}I_2,\ 10^{-6}I_2$$

where

$$r = \begin{pmatrix} 0.18 \\ 0.06 \end{pmatrix},\ R = \begin{pmatrix} 0, 1, 0 \\ 0, 0, 1 \end{pmatrix},\ I_2 = \begin{pmatrix} 1, 0 \\ 0, 1 \end{pmatrix}$$

The set of values of ψ describe a sensitivity test of the posterior estimates of β that is cognate to the Ridge trace. Robins and Greenland (1986), "... recommend that one always perform sensitivity analyses. That is, one should determine if moderate changes in one's prior beliefs or one's analytic procedures (such as one's model selection strategy) would lead to large changes in one's inferences about the parameter of interest. With weak non-experimental data, strong dependencies of inferences on prior beliefs will usually be found."

The ME estimate for $\psi = 10^{-6}I_2$ corresponds to the Supe et al (1983) estimate. It is unacceptable on two counts: 1) It lies outside the 0.90 confidence ellipse on α (Obenchain criterion); 2) The assumption $\psi = 10^{-6}I_2$ is inconsistent with the apriori (BIR) information.

The figure also shows the position of $\bar{\alpha}(4)$, the estimate of α obtained by deletion of observation at index number i=4. The obvious extreme lability, described by $(\bar{\alpha} - \bar{\alpha}(4))$, is due to the extreme degree of multicollinearity in the distribution of observations, as is shown in Fig. 4a.

estimates of the ratios $\theta_1 = \beta_2/\beta_1$ and $\theta_2 = \beta_3/\beta_1$ obtained from the more precise estimates of the parameter vector of the LQ-L model obtained from the BEIR III leukemia incidence data were used to stabilize the cognate sample estimates of the LQ-L model of the BEIR III non-leukemia cancer mortality data is described in Fig. 38. The parameter estimates of the LQ-L model of the leukemia incidence data are given in Table V-8 of the BEIR III Report; the Mixed estimates of the LQ-L model of the cancer mortality data are given in Table V-11 of that report. The deployment of this technique in the BEIR III report is questionable for several reasons of which two will be described briefly: 1) It represents an interspecies transfer of dose-response functions since leukemia and cancer are two different species of tumor with different natural histories and radiation sensitivities. (N.B. Peto (1977) divides malignant tumors into epithelial tumors (90%) and non-epithelial tumors (10%). The latter include the leukemias. The former can be further divided into the sex-specific (20%) and non-sex-specific (70%).) 2) The improvement in the precision of the parameter estimates of the LQ-L model of non-leukemia cancer mortality that are reported in Table V-11 of the BEIR III report could only be achieved by assuming a degree of precision for the a priori information represented in the ratios θ_1 and θ_2 that is inconsistent with that of the estimates of β_1, β_2 and β_3 in the LQ-L model of leukemia incidence. The parameter estimates in Table V-11 are based on the assumption that $\psi \equiv [0]$. But $\psi = RVar(\hat{\beta})R^T$ where the diagonal elements of $Var(\hat{\beta})$ are described in Table V-8 of the BEIR III report; it is quite evident from Table V-8 that the assumption $\psi \equiv [0]$ vastly overstates the precision with which the ratios θ_1 and θ_2 are known.

Moreover, it can be shown that the BEIR III estimates of the parameters of the LQ-L and Q-L models of the LSS(T65D) data on non-leukemia cancer mortality are dominated by a single extreme observation at $D_\gamma = 258$ rad, $D_n = 2.31$ rad. For the LQ-L model this observation (row #16) has Cook's D = 2.26; for Q-L Cook's D = 1.36. (But for the L-L model the parameter estimates are not dominated by this or any other single observation - or group of observations.) Deletion of the observation for which Cook's distance is $D_{16} = 2.26$ changes the sample estimates of the LQ-L coefficients of D_γ, D_γ^2, and D_n by factors of 2.31, 12.53, and 1.38, respectively. (N.B. It will be recalled that the sample estimates of the parameters of the BEIR III LQ-L model of the LSS(T65D) leukemia incidence data are dominated by a single non-extreme observation at $D_\gamma = 38.8$ rad, $D_n = 0.1$ rad. For the models of the mortality data the dominant observation can be identified by a large value of the hat matrix diagonal; for the model of the leukemia data the dominant observation can be identified by a large - albeit not very large - value of the residual. See Fig. 24.) However, Fig. 38c vividly discloses that the constrained estimates of the parameters of the BEIR III LQ-L model of non-leukemia cancer mortality (Table V-11 of the BEIR III report) are dominated by the two pseudo-observations that represent the information on dose-response of leukemia (See Table V-8) that is implemented by the BEIR III constraint. The consequent increase in the bias of these parameter estimates is the trade-off for the stabilization of the parameter estimates of the BEIR III models of the LSS(T65D) cancer mortality data that is reported in Table V-11 of the BEIR III (1980) report (Compare with Table V-9 of that report).

Figures 35 through 38c either demonstrate or suggest how all three of these post-hoc salvage methods (Ridge regression, Mixed estimation, Data augmentation) might be deployed to obtain, from dose-response and isoeffect models of non-uniform, multicollinear, clinical data, more useful assessments of the respective roles of fractionation and protraction in modulating the biological effects of radiation dose in the treatment of cancer.

It is evidenct from Table V-8 of the BEIR III report that the LQ-L model over-fits the LSS(T65D) leukemia incidence data. Nonetheless, this was the model of choice in the BEIR III report (NAS/NRC, 1980). It was also the model of choice for the non-leukemia cancer mortality in that report although leukemia is a non-epithelial tumor and cancer is an epithelial tumor. The former comprise only (about) 10% of all tumors, the latter comprise (about) 90% of all tumors. It was demonstrated in the BEIR V report (NAS/NRC, 1990) that for the LSS(DS86) data only the leukemia mortality rate has a linear-quadratic dependence on radiation dose. All other tumors have a linear dependence on dose. Now Plato (in The Sophist) observed some little while ago that, "A

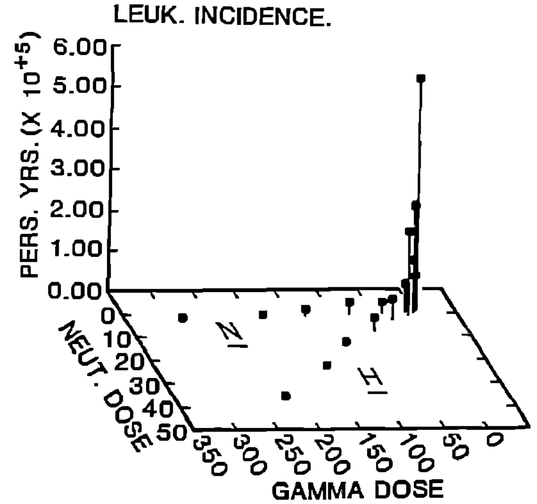


Fig. 37. Three-dimensional plot of person years at risk, nt, vs gamma dose, D_γ and neutron dose, D_n, for the leukemia incidence rate for the pooled city-specific samples: H-Hiroshima, N-Nagasaki. The large correlation of D_γ and D_n within each city is characteristic of the respective radiation sources: The Fat Man and Little Boy A-bombs. The large values of nt at the lower doses - the largest value is at $D_\gamma = D_n = 0$ - is a result of the spherical symmetry of the "experimental design": The intensity of the unfiltered radiations at each city can be described, approximately as, $I = I_0 r^{-2} e^{-kr}$ where k^{-1} is a characteristic relaxation length. The disease experience of all persons within an annulus with center at the hypocenter is pooled to provide an estimate of response at the average dose within the annulus. As distance, r, from the hypocenter is increased the average dose decreases and the area of the annulus, and hence the number of persons at risk, increases. (Reproduced with permission from Herbert, 1986b).

Because of the large multicollinearity in the distributions of the gamma, D_γ, and neutron, D_n, doses within the sample from each city the city-specific estimates, $\hat{\underline{\beta}}$ and $V\hat{a}r(\hat{\underline{\beta}})$, respective parameter vectors, $\underline{\beta}$, for the BEIR III (1980) LQ-L, L-L, and Q-L models of radioleukemogenesis are inflated and several of the estimates, $\hat{\beta}_j$, have incorrect signs. Therefore, the city-specific data were pooled, and estimates of the parameter vectors of the respective models were constructed from the pooled data. The Fig. describes a successful example of the salvage maneuver of Data Augmentation:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \underline{\beta} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

Here the matrices X_1 and X_2 have complementary correlation structures.

This salvage maneuver should be compared with the radiation toxicity data in Figs. 35 and 36 in which the effects of multicollinearity in X on the sample estimate of $\underline{\beta}$ are reduced by the post-hoc maneuvers of Mixed estimation and Ridge regression and with Figs. 34b and 34c in which the effects of multicollinearity are avoided by a pre-hoc maneuver: an orthogonal experimental design. Note that both the pre-hoc and post-hoc maneuvers depend upon apriori, or non-sample, information on $\underline{\beta}$.

It is obvious that a similar salvage maneuver - data augmentation - that is, pooling two sets of radiation toxicity data in each of which N and T are highly correlated, but with different correlation structures, in the respective joint distributions, could be successfully deployed to obtain useful information on the respective roles of fractionation and protraction in modulating the biological effects of radiation dose in studies in which the response of interest has a Binomial distribution - instead of a Poisson distribution as is the case for the LSS radiation carcinogenesis data. This possibility suggests the fruitfulness of the generalized linear model of radiation dose-response.

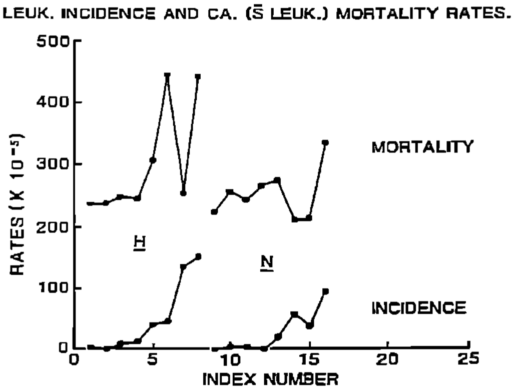


Fig. 38a. The figure presents an index plot of the leukemia incidence and cancer (sans leukemia) mortality rates for the BEIR III (1980) LSS data where H and N denote the Hiroshima and Nagasaki samples, respectively.

Figure 37 described the pooling of the Hiroshima and Nagasaki samples of leukemia incidence data. This salvage maneuver (Data Augmentation) was successful in achieving a lower degree of correlation between D_γ and D_n in the pooled sample than in the respective city-specific samples because the correlation structures of the latter are complementary (Herbert, 1986c). More acceptable sample estimates of the parameter vector $\underline{\beta}$ of the LQ-L, L-L, and Q-L models of leukemia incidence were obtained from the pooled data.

This maneiver also reduced the degree of correlation of D_γ and D_n in the cancer (sans leukemia) mortality data, of course. However, more acceptable sample estimates of the parameter vector, $\underline{\beta}$, of the dose-response models LQ-L, L-L, and Q-L still could not be obtained from those data. This is due to the fact that, since the response has a (conditional) Poisson distribution for which the variance is equal to the expectation, the (higher) mortality rates are noisier than are the (lower) leukemia incidence rates.

Therefore, the linear constraints, $\underline{r} = R\underline{\beta} + \underline{v}$, $E(v) = \underline{0}$, $Var(v) = \chi$, were imposed on the sample estimates of $\underline{\beta}$ for these models of the pooled mortality data in order to increase the precision of the sample estimate of $\underline{\beta}$. The elements of the matrices, $\underline{r}$ and R were obtained from the point estimates of $\underline{\beta}$ for the cognate models of the LSS leukemia incidence data. However, it was stipulated by the BEIR III Committee that $\psi = [0]$, the null matrix (although it is evident from Table V-8 of the BEIR III (1980) report that this stipulation greatly over-states the precision of the (pooled) LSS sample estimates of $\underline{\beta}$) for those models of the leukemia incidence data. (For example, in Table V-8 the sample estimates of β_1 and β_2 barely exceed the respective standard errors.)

The BEIR III estimates of $\underline{\beta}$ for the LQ-L, L-L, and Q-L models of cancer sans leukemia mortality were obtained as posterior estimates by using the procedure of Mixed estimation. For a Normal theory model we have

$$\begin{pmatrix}\underline{y}\\ \underline{r}\end{pmatrix} = \begin{pmatrix}X\\ R\end{pmatrix} \underline{\beta} + \begin{pmatrix}\underline{e}\\ \underline{v}\end{pmatrix}$$

This can be generalized to the Poisson models by the transformations,

$$P\underline{y} = PX\underline{\beta} + P\underline{e}$$

and

$$Q\underline{r} = QR\underline{\beta} + Q\underline{e}$$

where $PP^T = V^{-1}$ and $QQ^T = \psi^{-1}$ where $V^{-1} = W$ is the (n*n) Poisson weight matrix of the sample data. The proportion of apriori, or non-sample, information in the posterior estimates of $\underline{\beta}$ in Table V-11 for cancer (sans leukemia) mortality that is contributed by the leukemia incidence data of Table V-8 can be shown to be θ_p = 0.41 (See Herbert, 1986b, 1989d).

It is obvious that a similar salvage maneuver - Mixed estimation - can be deployed to obtain more precise estimates of the parameter vector $\underline{\beta}$ for models of radiation toxicity in which the response has a Binomial distribution just as in the present case in which the response has a Poisson distribution.

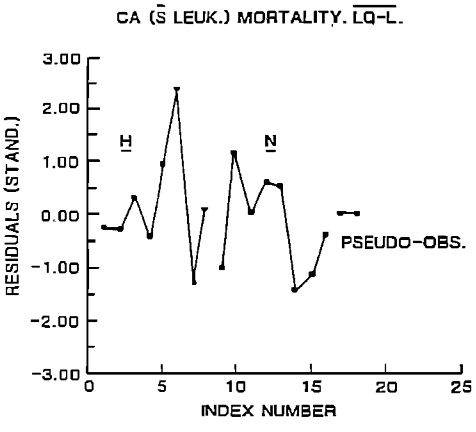


Fig. 38b. The figure presents an index plot of the standardized residuals for the LQ-L model of the BEIR III LSS (1980) cancer (sans leukemia) mortality data obtained by Mixed estimation:

$$\begin{pmatrix}P\underline{y}\\ Q\underline{r}\end{pmatrix} = \begin{pmatrix}PX\\ QR\end{pmatrix} \underline{\beta} + \begin{pmatrix}P\underline{e}\\ Q\underline{v}\end{pmatrix}$$

The two pseudo-observations at i = 17, 18 represent the leukemia incidence information on the LQ-L model that is conveyed by the constraint $\underline{r} = R\underline{\beta} + \underline{v}$; $E(v) = \underline{0}$, $Var(\underline{v}) = \psi = [0]$, where $\underline{r}$ is (2x1), R is (2xk), etc. The matrices P and Q are weight matrices appropriate to the cancer mortality (sans leukemia) sample data and the leukemia constraint, respectively: $P^TP = V^{-1}$ where V^{-1} is the (nxn) Poisson weight matrix of the sample obtained from the IRLS estimates of $\underline{\beta}$ for the LQ-L model of non-leukemia cancer mortality and $Q^TQ = \psi^{-1}$ is the (2x2) weight matrix of the leukemia constraint. See Belsley et al, 1980 and Herbert, 1986b, 1989d.

Note that the two pseudo-observations, which represent the leukemia incidence sample, dominate the estimates of $\underline{\beta}$ for the LQ-L model of radiation carcinogenesis (sans leukemia). This plot should be compared with those for the LQ models of the

Tradescantia radiation mutagenesis data in Figs. 5b, 6b, and 8 in which the observations in the sample of experiment #6 dominate the estimate of $\underline{\beta}$ in the pooled sample which includes the data of experiments #2, #5, #6, and #7.

It should be remarked that the use of apriori information on $\underline{\beta}$ that is obtained from the leukemia incidence data to stabilize the parameter estimates of the LQ-L, etc., model of cancer (sans leukemia) mortality data represents an instance of the "interspecies transfer of dose-response functions". But, "While the mechanisms of induction of most carcinomas may all be rather similar, they probably differ fundamentally from the mechanisms of induction of leukemias, sarcomas, etc., and much of the research into leukemia viruses, or sarcoma viruses may be irrelevant to the 90% of human tumors that are carcinomas" (Peto, 1977). DuMouchel and Harris (1983) have proposed, "... a class of Bayesian statistical methods for interspecies extrapolation of dose-response functions [Bayesian hierarchical meta-analysis]. The methods distinguish formally between the conventional sampling error within each dose-response experiment and a novel error of uncertain relevance between experiments ... the dose-response data from many substances and species are used to estimate the inter-experimental error. The data, the estimated error of interspecies extrapolation, and prior biological information on the relations between species or between substances each contribute to the posterior densities of human dose-response." ... "Human cancer risk assessment requires data on many agents in many species. In the absence of strong prior information on cancer mechanisms, one good rat study is just not enough."

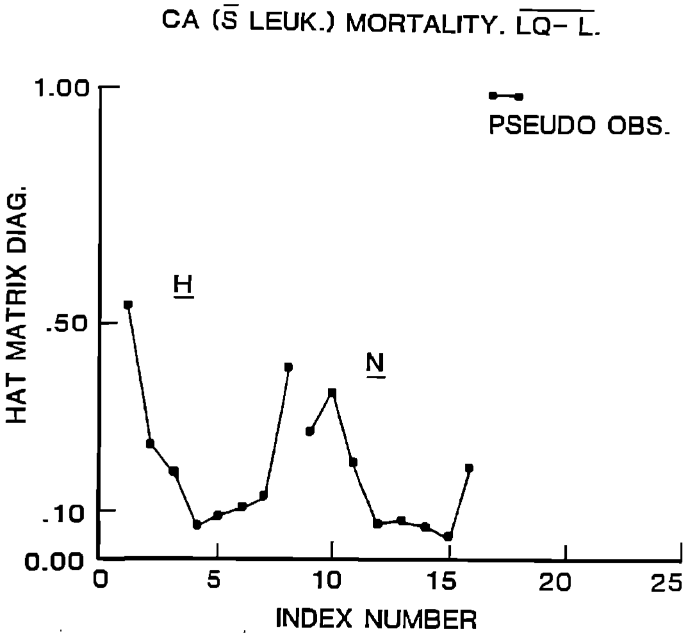


Fig. 38c. The figure presents an index plot of the hat matrix diagonals for the LQ-L model of the BEIR III (1980) LSS cancer (sans leukemia) mortality data obtained by Mixed estimation.

It is obvious that the two pseudo-observations representing the leukemia incidence information completely determine two of the five components, β_j, of the parameter vector, $\underline{\beta}$. This information from the diagnostics is consistent with the estimate $\hat{\theta}_p = 0.41$ obtained from the Mixed estimation technique. See Belsley et al, 1980 and Herbert, 1986b, 1989d.

Moreover, the other single-row deletion diagnostics DFBETAS, COVRATIO and DFFITS, as well as PRESS, are functions of the two key diagnostics, e_i^* and h_i. In general, these diagnostics will be large when either e_i^* or h_i - or both - are large. Thus, it appears that the sample estimates, $\hat{y}$, $\hat{\underline{\beta}}$, Var($\hat{\underline{\beta}}$) and PRESS for the LQ-L model of non-leukemia cancer mortality are dominated by the pseudo-observations #17 and #18 which represent the information on $\underline{\beta}$ obtained from the sample estimates of the leukemia incidence rate. This would seem to raise some issues of interpretability of the posterior estimate $\bar{\underline{\beta}}^{**}$ of the parameter vector $\underline{\beta}$ in the LQ-L model on non-leukemia cancer mortality.

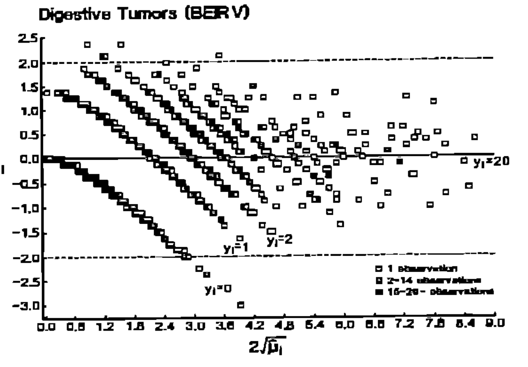


Fig. 38d. Plot of the Freeman-Tukey residuals, *gi* vs $\bar{\mu}_i$, for the preferred linear (in radiation dose) BEIR V (1990) model of the mortality rate for digestive tumors in the LSS sample (DS86 dosimetry). Note that $\bar{\mu}_i$ is plotted in the constant information scale, $2\sqrt{\bar{\mu}_i}$, for a Poisson error variable. The Freeman-Tukey residuals, $g_i = \sqrt{y_i} + \sqrt{y_{i+1}} - 4\sqrt{\hat{\mu}_{i+1}}$, are preferred over the deviance, d_i, or Pearson, χ_i, residuals for sparse Poisson data. Note that the first few contours of constant y_i have (approximately) a slope of -1 as is appropriate for a residual vs $\bar{\mu}_i$ plot when $\bar{\mu}_i$ is plotted in a constant information scale.

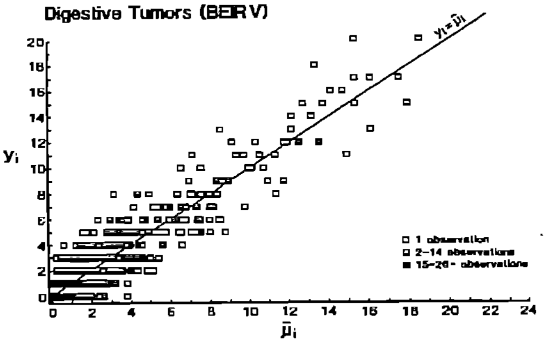


Fig. 38e. Plot of the observed, y_i, vs expected, $\bar{\mu}_i$, response for the BEIR V (1990) model of the mortality rate of digestive tumours in the LSS (DS86) sample.

cautious man should above all be on his guard against resemblances; they are a very slippery sort of thing." The findings of the BEIR V report suggest that he knew what he was talking about. Figures 38d and 38e are plots of the Freeman-Tukey residuals and observed responses vs the response, $\hat{\mu}_i$, estimated from the linear model for the digestive tumors in the LSS(DS86) data.

8.0. Further criteria for "believability". The between-sample invariance of the form and/or parameters of a regression model.

"The recognition of a subset of situations in which a degree of invariance occurs appears to be a crucial step in the development of a successful model."

J. Nelder, 1968

The presence of "pattern" in the set of residuals $\underline{e}$, of a specified model "fitted" to a single set of data identifies an inadequate model on the criterion of concordance. However, the presence of "pattern" in the parameter estimates $\hat{\underline{\beta}}$, of a specified model fit to several, say m, sets of different data identifies an adequate, or better, a "law-like" model, on the criterion of invariance. Nelder (1968) has put it very well: "What is to be the criterion for success in model-building? I shall adopt the proposition that we are succeeding to the extent that we can discern a pattern in the estimates of the parameters of the model, and that this implies the recognition of invariance. If we have several groups of data to which we fit straight lines and if it appears that [a] single slope parameter suffices for all, then we have established a degree of invariance over at least those sets of data." Ehrenberg (1975) has remarked that invariance of the parameter estimates of a model between data sets is a characteristic of law-like relationships: "A relationship becomes lawlike when different sets of data are summarized or modelled by the same quantitative equation. Its status depends upon the range of empirical conditions under which it holds." And Box et al (1978) have remarked, "Models that are inadequate for a given purpose do not necessarily show their inadequacy with a particular set of data ... . One device, useful not only in revealing in adequacy but also in pointing to its possible cause, employs the principle that in an adequate model constants stay constant when variables are varied." And McCullagh and Nelder (1989) have noted that, "An important property of a model is its scope, i.e., the range of conditions over which it gives good predictions. Scope is hard to formalize, but easy to recognize, and intuitively it is clear that scope and parsimony are to some extent related. If a model is made to fit very closely to a particular set of data, it will not be able to encompass the inevitable changes that will be found to be necessary when another set of data relating to the same phenomenon is collected. Both scope and parsimony are related to parameter invariance, that is, to parameter values that either do not change as some external condition changes or that change in a predictable way."

The secondary analyses presented in Annex II, parts 3 and 5, and in Annex III, part 6, describe several models in which the variance of a slope parameter between different sets of data is demonstrated and which therefore suggest that these models are, to this degree, "law-like" and might therefore be extrapolated with somewhat more confidence than, say a Taylor series model of the same process.

For example, it is shown in Annex II, part 3 that the parameter estimate $\hat{\beta}_1$ of M1a, the NSD-type model of the radiation toxicity data of Fig. 9 above is invariant between seven data sets characterized by different values of (N,T):

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = \beta_0^* + \beta_1 x_1$$

The intercept $\beta_0^* = \beta_0 + \beta_2 x_2 + \beta_3 x_3$ is a function of (N,T). The slope β_1 is invariant across the set of 7 experiments. In this model z is the probit transform, $z = \Phi^{-1}(\pi)$, where Φ is the Normal distribution function, $0 \leq \pi \leq 1$, and $x_1 = \log D$, $x_2 = \log N$, $x_3 = \log T$.

The parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the Target theory model of radiation lethality (cell survival), $m = \beta_0[1-(1-e^{\beta_1 D})^{\beta_2}]$ are invariant between the mouse and rat bone marrow stem cell survival data of Annex II, part 5. β_1 is the "slope parameter" and β_2 is the "intercept parameter" on the usual semi-log display of cell-survival curves.

Two further instances of the between-sample invariance of the estimates of the parameter vector of a dose-response model can be found in our secondary analyses of the Shellabarger et al (1969) and Montour et al (1977) data on radiation-induced mammary neoplasia in the Sprague-Dawley female rat in Annex III, part 6.

1) The sample estimate of the parameter vector $\underline{\beta}^T = (\beta_0, \beta_1)$ of the probit model, $z = \beta_0 + \beta_1 x_1$ of radiation-induced mammary neoplasia in female Sprague-Dawley rats (Shellabarger et al 1969) is invariant under deletion of the high dose observation at 500R (Co^{60}_{γ}). This suggests that for these data the response is pure induction of neoplasia, unattenuated by "cell-killing" at high dose, contrary to what has been asserted by others. See Fig.30.

2) The estimate of the slope parameter β_1 of the probit model, $z = \beta_0 + \beta_1 x_1$, of radiation-induced mammary neoplasia in female Sprague-Dawley rats, is invariant between two different data sets characterized by different values of the LET of the radiation x-rays and neutrons. Therefore, the pooled data of Shellabarger et al (1969) on γ-ray exposures of the Sprague-Dawley rat and the Montour et al (1977) on neutron exposures of the same species can be well-graduated by the model $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where x_1 = log dose and x_2 is an indicator variable for LET: $x_2 = 0$ (neutron), $x_2 = 1$ (X-ray). This suggests that for these data the neutron RBE is independent of dose, contrary to previous assertions. In this model the neutron RBE is estimated as RBE = 10^{θ} where $\hat{\theta} = \hat{\beta}_2/\hat{\beta}_1$. See Fig. 32.

The reader should note that in each of these several examples the demonstrated invariance - "law-like" behaviour - was characteristic of the rival model, but not of the received model with which it was compared, on common sets of data. In Annex II, part 3, the received model is the multifraction LQ model; in Annex II, part 5, the received model is the single-fraction LQ model; in Annex III, part 6, the received model is the linear model, $\pi = \alpha_0 + \alpha_1 D$.

Note, however, that invariance of the decrement of logS, or increment of damage, under permutation in the sequence of N fractions of size d - and hence between different levels of total dose (and damage) - is one of two fundamental postulates of the multi-fraction LQ model. See Thames et al, 1982; Fowler, 1989. But this invariance, unlike the invariance of the slope parameter β_1 of model M1a over the seven treatment regimens (N_i, T_i), $1 \leq i \leq 7$, each with a different range of dose, as described above, has not been demonstrated. Moreover, this postulate leads to the form $\eta = \alpha D + \beta dD$ for the linear predictor of the multifraction LQ model, which, since it includes the interaction term dD but not both of the main effects terms, d and D, is inherently implausible (Nelder, 1977). Moreover, it is inconsistent with a well-known biological principle of wide applicability which requires that the response of a biological system change in level and/or in nature under repeated exposure to the same insult or injury.

The reader should recall that the invariance of the sample estimate $\hat{\underline{\beta}}$ to small perturbations of the observation matrix $[\underline{y}, X]$ such as row-deletion is also a good criterion of the adequacy of the model. That is, the set $\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}}\ ^{-}\ \underline{0}$, $1 \leq i \leq n$, where $\underline{0}$ is the null vector and $\hat{\underline{\beta}}_{(i)}$ is the row-deleted estimate of $\underline{\beta}$, identifies a statistically adequate model. This raw measure of parameter lability is usually standardized by $Var(\hat{\underline{\beta}})$ to give Cook's distance, D_i:

$$D_i = (\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})^T[Var(\hat{\underline{\beta}})]^{-1}(\hat{\underline{\beta}} - \hat{\beta}_{(i)}).$$

Similarly, if the estimate of $\underline{\beta}$ is invariant under small perturbations of the covariance matrix, $Var(\hat{\underline{\beta}})$, then the model, $\underline{y} = X\underline{\beta} + \underline{e}$ is adequate. This can be seen at once in the Ridge regression estimates, $\bar{\underline{\beta}}_R = [X_0^TX_0 + kI]^{-1}(X_0^TX_0)\bar{\underline{\beta}}$, $0 \leq k \leq \infty$, in the correlation basis ($X_0^TX_0$ is a correlation matrix). $\underline{\beta}_R$ and $\hat{\underline{\beta}}$ are the Ridge and Least Squares estimates of $\underline{\beta}$ in that basis. The perturbation is described by the bias parameter k. Note that when $k = 0$, $\hat{\underline{\beta}}_R \equiv \hat{\underline{\beta}}$. Therefore, the criterion $[\hat{\underline{\beta}}_R - \hat{\underline{\beta}}]\ ^{-}\ \underline{0}$ for $0 \leq k \leq 10^{-1}$, identifies an adequate model of the sample.

9. Further comments on model validation.

"Validation: A procedure which provides, by reference to independent sources, evidence that an inquiry is free from bias or otherwise conforms to its declared purpose. In statistics it is usually applied to a sample investigation with the object of showing that the sample is reasonably representative of the population and that the information collected is accurate."

M. Kendall and W. Buckland, 1971

"The word 'valid' would be better dropped from the statistical vocabulary. The only real validation of a statistical analysis, or of any scientific enquiry, is confirmation by independent observations."

F. Anscombe, 1967

"Users have often been disappointed by ... multiple regression equations that 'forecast' quite well for the data on which they were built. When tried on fresh data, the predictive power of these (equations) fell dismally."

F. Mosteller and J. Tukey, 1977

9.1 General.

We now present a brief discussion of several methods and criteria of model validation that we have exploited extensively in the secondary analyses of the studies listed in Table 1. These methods and criteria are described at greater length and their use is illustrated with the data from the studies of Table 1 in the Annexes II-IV to which the interested reader is referred.

Two of the more lucid and accessible discussions of the validation of regression models are presented by Montgomery and Peck, 1982 and by Snee, 1977 and we quote liberally from both references. The paper by Stone, 1974 should also be consulted.

"Model validation is directed toward determining if the model will function successfully in its intended operating environment" ... "Proper validation of a regression model should include a study of the coefficients to determine if their signs and magnitudes are reasonable. That is, can $\hat{\beta}_j$ be reasonably interpreted as an estimate of the effect of x_j? We should also investigate the stability of the regression coefficients. That is, are the $\hat{\beta}_j$ obtained from new samples likely to be similar to the current coefficients? Finally, validation requires that the model's prediction performance be investigated. Both interpolation and extrapolation models should be considered." (Snee, 1977).

"Methods to determine the validity of regression models include comparison of model predictions and coefficients with theory, collection of new data to check model predictions, comparison of results with theoretical model calculations and data-splitting or cross-validation in which a portion of the data is used to estimate the model coefficients and the remainder of the data is used to measure the prediction accuracy of the model." (Snee, 1977). Data-splitting, or cross-validation, is an effective method of model validation when it is not practical to collect new data. (Snee, 1977).

The importance of determining the degree to which an estimated model can be generalized should be remarked. A regression model that describes only one set of data is simply a profound description of a unique historical event. "Saying something twice does not make it any more true, but requiring a model to hold for two sets of different empirical data would at least serve to establish whether it could in fact generalize. (Many theories can hold for one selected set of data, but are they known to apply to a sufficiently wide range of circumstances to be of any practical or academic interest?)" (Ehrenberg, 1975). On the other hand a model for which both the form and parameter values are invariant over a wide range of data is said to be "law-like". (Ehrenberg, 1968).

In the evaluation of received models that is discussed in this report (see especially Annexes II-IV), we have exploited several of the methods and criteria of model validation. For example, as remarked above, we have determined that both the parameter vector $\underline{\beta}^T = (\beta_1, \beta_2)$ and the characteristic function of the parameter vector, $f(\underline{\beta}) = \beta_1^{-1}\ln\beta_2$, of the Target theory model of bone marrow stem cell survival data are invariant between the mouse and rat data, whereas the parameter vector $\underline{\beta}$ and the function $f(\underline{\beta}) = \beta_1/\beta_2$ for the LQ model of these data are not. We have shown that the sign of the curvature, k, in the low dose region ($0 \leq D \leq 2.0$ Gy) of the dose-response curve for the Target theory model of these data agrees with theory whereas that for the LQ model does not: $k_T < 0$ vs $k_{LQ} > 0$. We have shown that the slopes of the respective rival models of the radiation toxicity and radiation neoplasia data are invariant under changes in (N, T) of the irradiation schedule and LET of the radiation, respectively. We have shown that the predictions of the rival (probit) model of the (spontaneous) response at $D = 0$ for the radiation neoplasia data is consistent with both the sample data and previous experience, whereas those of the received (linear) model do not, And so forth.

9.2. Cross-validation.

In cross-validation, the data set, size n, is partitioned into two parts, a so-called estimation, or training set size n^*, and the validation set size $n-n^*$. An estimate, $\underline{\hat{\beta}}$, of the parameter vector of the model is constructed on the training set and the predictions of the model say, $\underline{x}_i^T\underline{\hat{\beta}}$, $n^* + 1 \leq i \leq n$, are compared with the observations, say y_i, in the validation set, $y_i - \underline{x}_i^T\underline{\hat{\beta}}$, $n^* + 1 \leq i \leq n$. A squared loss criterion may be used to evaluate the model. The method of partitioning may be random, or it may be systematic, as in the "optimal" partition produced by the so-called DUPLEX method. (Snee, 1977). Moreover, the two partitions need not include equal numbers, i.e., $n^* \neq n/2$. However, as a rule of thumb, it is recommended that one not consider data-splitting unless $n > 2k + 25$ where k is the number of parameters to be estimated. "In order to have adequate degrees of freedom for error, giving reasonable power for significance tests and a meaningful residual analysis, the size of the estimation set, _" n^* should exceed $k + 10$. Moreover, "_. fifteen to twenty data points are required to obtain a reliable estimate of the prediction standard deviation."

Validation of dose-response models by comparison of the estimated parameter vector, $\underline{\hat{\beta}}$, or of the estimated response $\underline{x}_i^T\underline{\hat{\beta}}$, with the cognates obtained, a priori, from introspection, from theoretical considerations, or from other studies either experimental or epidemiological, that is, by examining the consistency of the sample estimates with non-sample information is necessary but not sufficient. A direct assessment of the predictive performance of the model in new data is required. However, new data are usually not immediately available and the size of the original sample from which the estimate $\underline{\hat{\beta}}$ was obtained is usually too small for data-splitting to be a viable alternative. In these circumstances we must consider still another measure of the susceptibility of the model to fruitful generalization: the PRESS statistic. (Montgomery and Peck, 1982).

9.3 Jackknife validation. The PRESS statistic.

In the evaluation of regression models it is becoming a common practice to leave out one data point at a time fitting the model to the remaining points, and predicting the response at the omitted point. The sum of the squared prediction errors, each point being left out once, is a cross-validation measure of prediction error, the PRESS statistic, which was briefly described under section 7.1 above. This statistic is sufficiently important, however, to merit some further discussion. In part, its importance derives from the fact that, like the Akaike Information Criterion (AIC), (Akaike, 1974) the PRESS statistic can be exploited to discriminate between non-nested rivals. (Stone, 1976). The PRESS statistic can also be used to provide cross-validation of a model when the original sample is too small to split into an estimation sub-sample and a validation sub-sample. (Stone, 1974).

The PRESS statistic provides a simple intuitive measure of the degradation of predictive performance of an estimated regression model in "new data", i.e., data different from that which yielded the sample estimate $\underline{\hat{\beta}}$ of the parameter vector $\underline{\beta}$. Thus, PRESS is a measure of the "regression" of the regression model. "A sensible estimate of the average prediction error, assuming that the data at hand are a sample from the population of future values, is then $(\text{PRESS}/n)^{1/2}$." (Allen, 1971, 1974).

The PRESS statistic, "... corresponds to division of the sample (size n) into a 'construction' subsample (size n-1) and a 'validation' subsample (size 1) in all (n) possible ways." (Stone, 1976). The PRESS statistic is a sum of squares of predictive residuals, $e_{(i)} = y_i - x_i^T\underline{\hat{\beta}}_{(i)}$ and thus is similar to the RSS statistic which is a sum of squares of ordinary residuals, $e_i = y_i - \underline{x}_i^T\underline{\hat{\beta}}$:

$$\text{PRESS} = \Sigma e_{(i)}^2 \text{ vs RSS} = \Sigma e_i^2$$

$\underline{\hat{\beta}}_{(i)}$, the row-deleted estimate of $\underline{\beta}$, is defined as $\underline{\hat{\beta}}_{(i)} = \underline{\hat{\beta}} - e_i(X^TX)^{-1}\underline{x}_i/(1-h_i)$

The residuals $e_i = y_i - \underline{x}_i^T\underline{\hat{\beta}}$ are based on a fit of the model to all the data. The i^{th} predicted residual $e_{(i)} = y_i - \underline{x}_i^T\underline{\hat{\beta}}_{(i)}$ is based on a fit to the data with the i^{th} observation (y_i,

$\underline{x}_i^T$) deleted.

The cross-validation estimate of expected excess error is $\Delta = n^{-1}\Sigma e_{(i)}^2 - n^{-1}\Sigma e_i^2$, the difference in observed error when we do or don't let x_i assist in its own prediction (Gilchrist, 1984); it is (PRESS-RSS)/n.

PRESS differs from the RSS in that no sampling distribution is defined for the PRESS statistic. The predictive residual, $e_{(i)}$ is a simple function of the residual ei:

$$e_{(i)} = y_i - \hat{y}_{(i)} = y_i - \underline{x}_i^T\hat{\underline{\beta}}_{(i)} = e_i/(1-h_i).$$

In using PRESS to discriminate between rival models the rule is to choose the model with the smallest value of PRESS. But, "Using PRESS as a criterion for model selection will result in a preference for models that fit relatively well at remote rows of X. To correct for this, Studentized versions of the predicted residuals and of PRESS can be suggested." (Cook and Weisberg, 1982). If the predicted residual is $e_{(i)} = e_i/\sigma\sqrt{1-h_i}$, then PRESS = $\Sigma e_i^2/\sigma^2(1-h_i)$. This version is similar to the weighted jackknife of Hinkley (Hinkley, 1977) described elsewhere (the "weight" in each case is a function of $(1-h_i)$).

Note that, "This form of cross-validation [PRESS] looks like the jackknife in the sense that the data points are sequentially deleted one at a time. However, there is no obvious statistic being jackknifed, and any deeper connection between the two data ideas has been firmly denied in the literature." (Gilchrist, 1984). These remarks notwithstanding, PRESS is frequently referred to as "jackknife validation". But it is the case, of course, that the PRESS statistic is a reduced-bias estimate of RSS = Σe_i^2 which is a biased estimate of the predictive performance of the model in new data: PRESS > RSS. Let us be more explicit. The jackknife methods of Hinkley (1977) provide a reduced bias estimate of the non-linear function of the parameter vector $\underline{\beta}$, namely, $\hat{\theta} = f(\hat{\underline{\beta}}) = \hat{\beta}_1/\hat{\beta}_2$, in terms of the row-deleted estimates $\hat{\underline{\beta}}_{(i)}$ with respective weights $(1-h_i)$. The jackknife methods of Allen (1971, 1974) provide a reduced-bias estimate of a linear function of the parameter vector $\underline{\beta}$, namely, the predictive performance in new data, $\Sigma(y_i^* - \underline{x}_i^{*T}\hat{\underline{\beta}})^2$, in terms of the row-deleted estimates $\hat{\underline{\beta}}_{(i)}$, $\Sigma(y_i - \underline{x}_i^T\hat{\underline{\beta}}_{(i)})^2 = \Sigma e_{(i)}^2$. (The $(y_i^*, \underline{x}_i^{*T})$, $1 \leq i \leq n^*$ identify the new sample.)

Note that the comparison of the respective PRESS statistics for the models M_j and M_k will provide one answer, albeit only a partial one - an "appearance of an answer" - to the initial question raised by the Task Group as to the nature and size of the losses l_{jk} incurred if the received model, say M_j, is deployed when the rival model, say M_k, is correct. (See section 1.4 above.)

It is of interest to note that two of the methods used to validate a dose-response model may, in certain circumstances, also be used to obtain improved estimates of the parameters of the model - or of a function $\theta = f(\underline{\beta})$ of the parameters. Thus, comparison of $\hat{\underline{\beta}}$ with a priori information on $\underline{\beta}$ becomes Mixed estimation (and similarly for comparison with a priori information on $\underline{x}^T\hat{\underline{\beta}}$. (Allen and Jordan, 1982).) And, jackknife validation, which exploits the set of row-deleted residuals, $[e_{(i)}]$ to obtain a reduced-bias estimate of concordance of the model with new data, measured by PRESS becomes jackknife estimation, which exploits the set of row-deleted parameter estimates $[\hat{\underline{\beta}}_{(i)}]$ to obtain a reduced-bias estimate of non-linear functions of $\underline{\beta}$, say $\hat{\theta} = f(\hat{\underline{\beta}})$. Note that in the construction of both PRESS and the weighted jackknife estimate of θ, the hat matrix diagonals appear as the weights, $(1-h_i)^{-1}$ and $(1-h_i)$, respectively.

9.4 Comparison of cross-validation and jackknife validation.

"Though a model's fit improves with the number of variables included, its predictive ability for new data does not necessarily improve. Given new, validation, data $y_{n+1}, ..., y_{n+v}$, with the corresponding regressor variable data, then a measure of model quality is obtained by using the model to obtain predictors $y_{n+1}, ..., y_{n+v}$. The natural criterion is the predictive sum of squares

$$\text{PRESS} = \sum_{i=1}^{v}(y_{n+i} - \bar{y}_{n+i})^2$$

The set of regressors leading to the smallest PRESS gives the best model.

Often such validating data does not exist so we make use of what can be called Jackknife Validation. We predict $\hat{y}_{i}$, within the data $y_1, ..., y_n$, by using the fitted model based on all the observations except the i^{th}, y_i, giving $\hat{y}_{(i)}$. For any set of regressors we can do this for all n observations separately and obtain

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2.$$

In each term $(y_i - \hat{y}_{(i)})^2$ we use a prediction of y_i based on 'other data'. PRESS will not necessarily decrease just by increasing the number of regressors, it gives a genuine measure of the quality of the model being used. Clearly the evaluation of PRESS for many variables and many observations involves a great deal of computation but it is a natural and effective criterion worthy of more extensive use" (Gilchrist, 1984).

As is the case with the other statistical concepts, methods, and criteria that are presented in this report, the uses of the PRESS statistic and the Akaike Information Criterion (AIC) in model discrimination are discussed and illustrated with data from the studies listed in Table 1 in Annexes II-IV.

Important Topics

Dose-response model. Isoeffect model. "Isoeffect transformation". Normal distribution, Binomial distribution. Poisson distribution. Weber-Fechner law. survival curve. "Survival transformation". Cell Inactivation Plot. Stochastic and non-stochastic effects. Goodness-of-fit statistics. Residuals (deviance, Freeman-Tukey, Pearson chi-squared). Nested models. AIC. Bayesian methods. Bayes factor, Prior odds ratio. Posterior odds ratio. Cook's D_i. trompe d'oeil measures of concordance. Invariance criteria. Data-instigated hypothesis. Methods-instigated hypothesis. Experimental design. Model validation. Cross validation. Jackknife validation.

10. Statistical methods and the ethics of scientific enquiry.

"The enormous amount of published research makes it inevitable that papers will often be judged, in the first instance at least, by the authors' own conclusions or summary. It is thus vitally important that these contain valid interpretations of the results of the study, since the publication of misleading conclusions may both nullify the research in question and falsely influence medical practice and further research."

D. G. Altman, 1982

There are, of course, insistent ethical, as well as scientific, issues addressed by a requirement for the deployment of appropriate statistical methods. Altman (1982) has discussed the issue: "So what is the relation between statistics and medical ethics?" ... "Stated simply, it is unethical to carry out bad scientific experiments. Statistical methods are one aspect of this. However praiseworthy a study may be from other points of view, if the statistical aspects are substandard then the research will be unethical. There are two principal reasons for this. Firstly, the most obvious way in which a study may be deemed unethical, whether on statistical or other grounds, is the misuse of patients (or animals) and other resources. "... one of the most serious ethical problems in clinical research is that of placing subjects at risk of injury, discomfort, or inconvenience in experiments where there are too few subjects for valid results, too many subjects for the point to be established, or an improperly designed random or double-blind procedure."

"Secondly, however, statistics affects the ethics in a much more specific way: it is unethical to publish results that are incorrect or misleading. Errors in the use of statistics may occur at all stages of an investigation, and one error can be sufficient to render the whole exercise useless. A study may have been perfectly conceived and executed, but if it is analyzed incorrectly then the consequences may be as serious as for a study that was fundamentally unsound throughout." Moreover, "If the results go unchallenged the researcher(s) involved may use the same substandard statistical methods again in subsequent work, and others may copy them. ..." (emphasis added). Seneca commented some while ago on such propagations of error: "No man errs for himself alone, but he is the cause and author of another's error, and error transmitted from one to another tosses and drives us headlong and we come to grief by the examples of other men" (quoted by R. Bacon in Opus Majus, 1271). For recent and relevant instances of investigators who have "... come to grief by the examples of other men," compare Figs. 2a, 10a, and 22, Figs. 13a and 13b, and Figs. 42a and 42b. See also Annexes II and III.

More recently, Marks, et al (1988) noted that, "Some observers have concluded that scientific experiments of poor quality are unethical. The conduct of the study whose design is fatally flawed and ultimately submitted for publication is really unethical. The investigator has wasted resources, peoples time, and put subjects at risk of experimentation. Basically, if two years of the researcher's effort goes down the drain because no valid conclusions evolve from his study then he has wasted two years of his life and countless hours and years of others people's lives. The really unethical medical studies conducted today are not those involving experimental drugs or those that lack Investigational Review boards scrutiny. They are those studies done without sufficient forethought at the outset, a waste of time and effort."

The pernicious effects of the regular failure of most investigators to present in their published reports statistically adequate measures of 1) the goodness-of-fit with their data and the consistency with generally received a priori beliefs of the dose-response model selected by the investigator to convey the information contained in the sample data (Recall the Robins and Greenland (1986) recommendation that the investigator "... choose a model that is consistent with the data and yields parameter estimates that are consistent with prior beliefs.") and 2) the precision of the sample estimates of the model parameters of a study, are exacerbated by the effects of yet another weakness that is present in virtually all of the published reports on radiation dose-response. This is the absence from the report of an accurate, precise, and circumstantial account of the primary data from which the findings that are reported were obtained. Obviously, if the full set of primary data were published in the report then the reader could determine for himself whether - and in what sense and to what degree - the dose-response model that was deployed by the author "fits" the data and whether the parameter estimates are precise. Moreover, if the model offered in the report appeared to be poorly chosen, i.e., if the forms of $f(\underline{x}_i, \underline{\beta})$ of the deterministic part and/or the distribution of the random part appeared to be misspecified (for example the deterministic part of the response may be either under- or over-fit, or a Normal (instead of a Binomial) distribution may have been chosen for the distribution of the random part of an observed binary response) - then the reader may be able to construct rival models of the same data that provide more defensible interpretations of what the data presented in the report may mean. Or, if the study is poorly designed, it may (but see below) be possible to salvage it, at least in part, by pooling the primary data of two or more similar studies (data augmentation) or by including a priori information on the parameters (as in Ridge regression or Mixed estimation, or on the response (Mixed estimation)).

Finally, the failure of most of the published reports in radiation dose-response to present either statistically adequate summaries of the findings or the primary data itself of the corresponding study precludes their salvage by the systematic integration of the research findings of two or more studies by means of a recently developed statistical procedure that has become known as meta-analysis (Glass et al, 1981; Halvorsen, 1986; DuMouchel and Harris, 1983). Meta-analysis is simply the application of statistical methods to the respective properties - primary data - of sets of two or more studies and/or to their statistical summaries - estimates and inferences - in order to achieve a quantitative integration, or synthesis, of the findings - rather than the more familiar narrative - and anecdotal - summaries of the findings of the respective studies; the latter, of course, being the common practice in those reviews of the literature that are currently published. As Light and Smith (1971) noted, concerning the usual (narrative) reviews that regularly appear in every literature, "Little headway can be made by pooling the words in the conclusions of a set of studies. Rather, progress will only come when we are able to pool, in a systematic manner, the original data from the studies."

A fruitful example of the salvage maneuver of pooling two sets of data with complementary weaknesses is provided by the LSS(T65D) leukemia incidence and non-leukemia cancer mortality data of the BEIR III (1980) report. (NAS/NRC, 1980). The city-specific data from Hiroshima and Nagasaki are weakened by the presence of the high correlation in the distributions of the gamma (D_γ) and neutron (D_n) doses so that useful city-specific dose-response models cannot be constructed. Pooling the data from the two cities reduces the correlation between D_γ and D_n and thus supports the construction of more useful Poisson linear models of these responses - including the cognate model of the LQ hypothesis. Figure 37 describes the pooled leukemia incidence data. Such pooling of non-experimental data is analogous to the data-augmentation salvage maneuver for flawed experiments.

The Tradescantia mutagenesis data obtained in the four experiments described by Sparrow et al (1972) and which was pooled in the NCRP 64 (1980) study in order to obtain estimates of α/β for the LQ model of this response provides, in contrast, a fruitless example of pooling data, that is, of data augmentation intended to correct weaknesses in the distribution of observations obtained in experimental studies, but in fact does not do so.

In the case of the radiation carcinogenesis data of BEIR III, the two sets of city-specific observations were mutually consistent and therefore the pooled data were homogeneous. However, in the case of the radiation mutagenesis data of NCRP 64, the four sets of experimental observations were mutually inconsistent and the pooled data were heterogeneous. Thus, the attempt to strengthen the parameter estimates of the model by meta-analysis was frustrated by the heterogeneity of the data ("Before combining results [or data] we must consider whether the results [or data] of different studies are homogeneous." K. Halvorsen, 1986). In each case, the tests of homogeneity were made by introducing indicator variables. For the carcinogenesis data, the variable is simply (0, 1) since there are only two possible sources for the data. For the radiation mutagenesis data the four experiments could be identified by the three indicator variables (0,0,1,0), (0,0,0,1) and (1,0,0,0). See Annex III, part 5.

It will be recalled that in several of the studies that we have reviewed the sample estimates of the α/β ratio of the LQ model were dominated by a single observation in the experiment (See Figs. 2 and 22) and in the BEIR III (1980) report both the sample evidence for the LQ-L model of leukemia incidence and the sample estimate of the cross-over dose (NIH, 1985) were dominated by a single anomalous observation. See Annex III, part 4 and Fig. 25a. Thus, it is of interest to note that for the mutagenesis response the sample estimate of the parameter vector, $\underline{\beta}$, of the LQ model of the pooled data is dominated, not by a single observation but by the group of five observations in one of the four experiments, thereby providing an example of Simpson's paradox. (Halvorsen, 1986; Simpson, 1951). See Figs. 5, 6, 7, and 8.

Several investigators have remarked that restriction to the integration of original data, and the methods based on it, might discard too many informative studies for which the primary data are inaccessible to the reviewer, therefore meta-analytic procedures have been developed to integrate summary statistics, such as parameter (say, the slope) estimates or p-values, as well as primary data (Glass et al, 1981).

But clearly the procedures of meta-analysis can be deployed to integrate summary statistics only if the studies to be integrated do report, or can provide by a secondary analysis, statistically adequate summaries of their findings. Indeed, the immediate goals of secondary analysis are either to verify that the study summaries that are reported are statistically adequate, or if they are not, and the primary data are available, to provide statistically adequate parallel analyses of a set of related studies since, "... it is useful to analyze data from the two [or more] studies using so far as is possible, identical methodology; and then to compare the findings. In this way any differences between the two studies will be highlighted, and if none are found the results will be in suitable form for combination" (S. Darby, 1986).

Of course, if the experiment is poorly designed, say there is multicollinearity in the X matrix or, especially, if the numbers at risk are quite small, say n_i = 5 or 6, for response with a Binomial distribution, as is now so often the case, then it may be impossible to retrieve much of value from the study by any re-analysis, salvage, or integration maneuver, whatever. The study may simply be a "one-off" report of a unique historical experience of the investigator from which it is impossible to generalize the findings. The published account simply awards an increment to the investigators' reputation without increasing the general fund of knowledge of the peer-group that makes the award.

However, the primary data are rarely published and they are usually not otherwise readily accessible to the reader. (A reader requesting the primary data of a study from the investigators is likely to experience a version of the so-called Wolins Effect[10] (Wolins, 1962). It is of interest to recall that Kepler constructed his famous laws on Tycho Brahe's data that he had <u>stolen</u> from Brahe's heirs.) In particular, for Binomial responses the numbers at risk, n_i at $\underline{x}_i^T$, and the numbers of responders, r_i out of n_i, are usually not published and, occasionally, when published, have been found to be badly misrepresented - by factors of as much as 2 or 3 For example, in Ang et al (1987), the number at risk at each level of dose can be read as n_i = 16. However, a check of the original data discloses that n_i = 5. It is worth remarking that we continue to find, as recently as 1990, <u>published</u> papers describing dose-response experiments in which the observed response has a Binomial distribution but which fail to include the numbers of animals at risk, n_i, in the report. In the case of one such paper, a subsequent request to the first author, readily provided these numbers: $4 < n_i < 15$. We recognize at once in these numbers two familiar weaknesses that we have remarked before in other studies: 1) The numbers at risk are too small. 2) The numbers at risk are non-uniform.

In two of the papers that we examined, the data have been found to consist of fictive observations, that is, <u>non-experiential</u> data. (See Annex II and Appendix I.) And although data on the dose, and the other covariates such as fractions and time, can occasionally be acquired by the reader from the artistically attractive graphs that are usually presented in the published papers, quite often this salvage maneuver, too, is frustrated by the investigators of the report, since the co-ordinates of the graphs are usually functions in which the primary observations on dose, and other covariates, are wrapped into such measures as (D/N) or TDF, that cannot be subsequently "unpacked" by the reader.

But failure to publish accurate primary data would seem to be a professionally culpable omission for, "One of the most widely held tenets of science is that research should be conducted <u>and reported</u> in a manner that yields sufficient information to enable people other than the original researchers to assess its merits and to replicate it" (T. Hedrick, 1985). Indeed, "Since the Baconian experimentalist scientists have insisted upon both accurate and circumstantial reporting." T. Kuhn, 1977. However, too many of the published reports on models of radiation dose-response are neither accurate nor circumstantial.

The failure of investigators either to take into account the prior information on the inherent variation of their observations, that is, the weights, w_i, $1 < i < n$, that is provided by the form of the distribution of the random part (Binomial or Poisson) of the observed response, or to adequately compare the predicted and observed responses ("goodness-of-fit"), or to publish their primary data, strongly suggests that, in the current praxis of radiation biology, the empirical data on an issue are not taken very seriously as evidence for an opinion. That is to say, the current investigations in the field of radiation biology appear to reflect an excessively Cartesian view of the proper practice of science.

One of the earliest published instances of an evident failure to take data very seriously is provided by two still frequently cited studies on the so-called "volume-effect" (von Essen, 1960, 1963). See Appendix I. In these studies the least squares estimates of the exponents of the volume factor are obtained by a regression on the <u>hypothetical</u> location parameters (intercepts) in families of (putative) volume-specific isoeffect curves ("In addition, only two of the isoeffect lines have been derived directly from the data. The construction of all the other curves parallel to these two originally derived curves is, then, hypothetical." (emphasis added)) Moreover, of the two "data-based" isoeffect curves referred to, one curve was "derived" from data on skin cancer response (ablation) and the other was "derived" from data on skin response (necrosis) by determining, "by eye", the intercept and slope of the respective discriminant curves for skin response (necrosis/no necrosis) and tumor response (ablation/recurrence). However, it is the case that such discriminant curves are - by definition - much closer to a $\pi = 0.50$ level of response (50% isoeffect) than to either of the levels of response reported in the paper: 3% skin necrosis, 99% tumor ablation. Thus, not only the intercept and slope but also the level of response, π, of the isoeffect curves in these volume-effects papers are "hypothetical". See Appendix I.

The evidence that data are not taken very seriously, e.g., the absence of adequate measures of the "fit" of model and data from those published reports that appear to have had a major impact in radiobiology and the nearly universal failure of any of the professional literature to include primary data in the published reports suggests that radiation biology is not necessarily practiced as an empirical science. Rather, it would seem that the results reported in the literature of this science - as it is now practiced - do not depend on "data" for their empirical validity. Indeed the results reported may even be confuted by the data adduced in their support - when compared with these data by statistically adequate methods. It is shown in Annexes I-IV that statistically adequate secondary analyses may often disclose that the reported results are inconsistent, in consequential way, with the data cited in the report. (It is also shown that the reported results may be inconsistent with the relevant theory as well as with the data cited.)

But, "Learning from experience", that is, from a correct, statistically adequate, analysis of empirical observations, is the method of empirical science. For, as Leamer (1970) has remarked, the proper function of data is to change opinions; that is, to map prior opinions into posterior opinions on the matters at issue. However, unless the goodness-of-fit of the model and data is assessed by statistically adequate measures (both aggregate and case statistics), neither the author(s), nor the reader(s) (including the "peer-reviewer(s)") of the study can possibly know whether the estimates and inferences that are based on it are truly warranted by the data or merely represent the results of either an inadequate analysis, or of believing what we want to believe. See remarks on the scientific method (Jeffreys, 1961). Thus, in the absence of statistically adequate evidence that the dose-response models selected by the author of a report do indeed "fit" the data of the report and yield parameter estimates that are consistent with prior information on the issue, the group of radiation scientists cannot either identify what it is that they do know, or recognize that which they do not know. In short, this distinguished group risks becoming intellectual Bourbons,[11] unable either to forget anything or to learn anything from the clinical, epidemiological, and laboratory experiences described by the data that are reported in the currently augmenting scientific literature on the important issue of radiation dose-response models. ("Models ... are what provide the group with preferred analogies or, when deeply held, with an ontology. At one extreme they are heuristic ... At the other they are the objects of metaphysical commitment." T. Kuhn, 1962.) Instead, there will continue to accumulate, by the relentless production of unevaluated papers on modelling, an inventory of equally "plausible", (that is, not inconsistent with some subset of prior beliefs - and purposes) several rival models, none of which can be said to be properly warranted by any empirical data - and hence none of which can be said to be "believable" ("plausible" maybe, but not "believable") - other than that it can be closely identified with some "believable" - perhaps charismatic - individual investigator or institution. (It may appear that this forecast differs sharply from that of Bailey (1967) which may be taken to represent the received view on the matter: "Provided that a constant flow of relevant experimentation is undertaken, most errors that are committed will be eliminated sooner or later by the inherent self-correcting mechanism of the scientific method." The apparent sharp difference is due to our recognition that in the absence of statistically adequate measures of "fit", the "inherent self-correcting mechanism" is broken - or at the least, badly bent.)

Moreover, although the purpose of any dose-response model is to relate, in a useful form, the variables of "dose" and "response" it is also necessary to know something both of those variables which are unimportant and of circumstances in which the model does not hold in order to deploy the model successfully. But this important information is also lacking in the absence of the deployment of statistically adequate measures of goodness-of-fit and the reporting of accurate and circumstantial accounts of the data. Indeed, because of these omissions it is often impossible even to discriminate between rival models of a given data set and thereby identify either as, say, more "believable" than the other, in the circumstances characterized by the data.

For example, until quite recently the multifraction LQ model of radiation toxicity did not include a time factor; in the initial parameterization of the LQ hypothesis the duration of irradiation, T, was an unimportant variable. Now, however, a time factor, $-\gamma T$, has been included, albeit in a rather ad hoc manner. (Travis and Tucker, 1987. A secondary analysis of this report is presented in section 16 of this report.) But the goodness- of-fit of the thus augmented LQ model has never been assessed by statistically adequate criteria either. Not to put too fine a point on it, for the received data that we have assessed as described in Annex II, the addition of the factor $-\gamma T$ to the LQ model did not improve the "fit" - the decrement in the residual sum of squares is not statistically significant (an instance of the so-called category 1 model check). See Annex II, part 3. However, models of rival hypotheses, with different parameterizations, which include a time factor, provided a better fit to the same set of data - on the evidence of both aggregate and case statistics. Thus, it seems to be the case that the LQ model is misspecified, perhaps badly so. That is, it is not only underfit - since the time factor is omitted - but the parameterization of the dose and fractions factors that are included are also misspecified. Actually, there is some additional prior information of a quite general sort that also suggests that the linear predictor, η, of the multifraction LQ model is misspecified. $\eta = \alpha D + \beta D^2/N$ obviously includes an interaction term $D(D/N) = Dd$ in dose per fraction, without the main effects term $d = D/N$. But, "... it is important that the final model or model make sense physically; usually this will mean that interactions should not be included without main effects nor higher degree polynomial terms without their lower-degree relatives" (McCullagh and Nelder, 1983). Breslow and Day (1980) have put the argument more forcefully - and in a biological context: "Models which contain interaction terms without the corresponding main effects terms correspond to hypotheses of no practical interest." This suggests that the hypothesis that gives the multifraction ($N \geq 1$) LQ linear predictor its characteristic form, namely $S_N = \Pi S_1$, where $S_1 = \exp(\alpha d + \beta d^2)$, has no interest for radiation oncology; the survival sequence, S_N, ends in a non-sequitur. Thus, neither the random nor the deterministic parts of the multifraction LQ hypothesis are correctly specified in most of the published studies in which they appear. See Annex II, part 3 and part 5, as well as sections 7.9.1 and 7.9.2 of this report.

11. C'est magnifique?

In 1857 General P. Bosquet, on observing the innovative deployment of Lord Raglan's light cavalry brigade at Balaclava, remarked, "C'est magnifique, mais ce n'est pas la guerre." A similar remark might seem to be a fair summary of much of the current untethered development of the ever more subtle, successively richer, and increasingly sophisticated models of radiation dose-response that are never adequately compared with the data for which they are offered as description and explanation: "C'est magnifique" - but it is not science. Or, as Kuhn (1970a) has remarked in another context (pre-Newtonian optics) "... though the field's practitioners were scientists the net result of their activity was something less than science."

11.1 Are there two sciences?

But perhaps the foregoing judgment is too facile - and it may seem to some, perhaps, a bit arch, as well. It may be lacking somewhat in both finesse and perspective. But, on the other hand, the common omissions (from the reports that we have examined), both of primary data and of any statistically adequate evaluations of the relationships of the proposed model to that data, or to related prior information on the matters at issue which would tend to complicate the received exposition (and which, it appears, post hoc, from the secondary evaluations of the studies that are described above, would often quite confute the investigators' explanation of what his data mean) that seem to distinguish the more influential published papers on dose-response models, do recall the perceptive remarks of the Polish pathologist/philosopher L. Fleck (Fleck, 1979) who discriminates between exoteric, or popular, science and esoteric science in the following ways: "Characteristic of the popular presentation is the omission both of detail and of controversial opinions; this produces an artificial simplification. Here is an artistically attractive, lively, and readable exposition with last, but not least, the apodictic valuation simply to accept or reject a certain point of view. Simplified, lucid, and apodictic science - these are the characteristics of exoteric knowledge. In place of the specific constraint of thought by any proof, which can be found only with great effort, a vivid picture is created through simplification and valuation." In the present context, the proof required to constrain thought - to distinguish the "useful model" (Box, 1979) from the "inspired blunder" (Koestler, 1965) - is, of course, provided by the goodness-of-fit statistics and the model checking procedures such as pattern recognition, regression diagnostics, etc., described above. Therefore, a more philosophically informed, as well as a more subtle, assessment of those selections of the current literature on radiation dose-response that were assessed for this report is that they appear to represent the achievements of an endeavor that Fleck has defined as exoteric science.

Thus, at least one of the "Two Cultures" identified by C. P. Snow in the nineteen-fifties, may be further subdivided, on the basis of rather similar criteria, into L. Fleck's "Two Sciences": The esoteric and the exoteric. We recall that Kuhn (1970a) has proposed a rather different dichotomy into normal and extraordinary science while Francis Bacon (1620) made a distinction between "Anticipations of Nature" (bad science) and "Interpretations of Nature" (good science). An alternative dichotomy of the current literature, which correctly reflects the exigencies of its times, divides it into reports by those who have to say something and reports by those who have something to say.

11.2 A partition of studies according to "quality".

The work described in this report suggests that the studies published in the literature may be usefully grouped according to their evidentiary, or probative, value on the issues that they address as follows:

A. Those studies for which the documentation in the published report is inadequate, i.e., those for which the reporting is neither accurate nor circumstantial, e.g., Tucker and Thames (1983); von Essen (1960, 1972). The only argument for the validity of the conclusions in such a report is that, because of the absence of any evidence, it is nearly impossible for the reader to demonstrate that they are not. (N.B.: The historian Aydelotte has remarked that, "The imprecise or slipshod formulation is impregnable; a statement that has no meaning cannot be disproved.")

B. Those studies for which the documentation in the report is adequate, that is, the primary data of the study are accessible and the primary analysis is reported in sufficient detail that it may be reproduced and the validity of the reported results of the study may be verified by the reader. (See also comments of Marks et al, 1988 in section 7.11.1.) Studies that are adequately documented in the published report can be further subdivided as follows according to the results of secondary analyses:

B1) Those studies in which both the analysis and the design are weak. These are wholly statistically inadequate studies, e.g., Thames et al, 1982, and all studies based on an F_e-plot. (Tucker and Thames, 1983 could also be included in B1) if one accepts the reported numbers at risk, $n_i = 10$, $1 \leq i \leq 19$, as valid.)

B2) Those studies in which the analysis, but not the design, is weak, e.g., Shellabarger et al, 1969; BEIR III, 1980. We note that although secondary analysis of studies of this type may salvage the study data, the findings from secondary analysis may well contradict the findings from the original, primary, analysis; it will not always "save the phenomena" that were described by the original authors (See, for example, Figs. 21, 30, and 32. Good data can, sometimes, triumph over bad analysis).

B3) Those studies in which the design, but not the analysis, is weak. No examples were encountered in the set of reports reviewed by the task group which are listed in Table 1. However, studies of this type can often be salvaged by data augmentation - or other forms of meta-analysis.

B4) Those statistically adequate studies for which neither the design nor the analysis is weak and therefore the inferences and estimates can be verified and accepted, e.g., Till and McCulloch, 1961; Frome and Beauchamp, 1968.

It is important to note that these latter judgments B1)-B4) can only be made in a responsible manner on the basis of a (statistically adequate) secondary analysis of the original data.

In the selections of the literature that were reviewed for this report the greatest number of studies fall into group B1) and the smallest into group B4). Our conclusions, based on these secondary analyses, are, regrettably, fairly consistent with those of other recent evaluations of the medical literature. For example, in an informal survey of the reports on radiation dose-response modelling that were published between 1980-1989 in a leading international journal we found that only 5 out of 86 (6%) reports were statistically adequate (suggesting that something less than 10% of the current reports in this field may warrant serious further study). Friedman (1988) in a recent editorial in JAMA has remarked that the state of statistical analysis in the medical literature is, "something of a disgrace". And in the recent review cited in the Introduction and summary, Williamson, Goldschmidt and Colton (1986) ask the question: "Are these findings [of the 28 assessments] generalizable? ... First, the assessments usually involved leading medical journals ... since the leading journals tend to be more selective than others, the quality of the entire medical literature is likely to be worse than these findings indicate ... Second, the publications samples revealed consistently poor quality; ..."And, "Overall, these findings may be typical for three reasons. First, poor scientific quality was found in nearly all the research - regardless of type or content ... Second, when design, analysis and documentation were assessed concurrently, the proportion of articles meeting the criteria often dropped to less than 1 per cent. Third, similar findings have been confirmed in the physical science literature. In our opinion, with respect to the clinical literature in English, the findings of the 28 assessments probably overstate the scientific quality of research publications in the applied health sciences." ... The findings reported in the 28 assessment articles suggest that the average practitioner will find relatively few journal articles that are scientifically sound in terms of reporting usable data and providing even moderately strong support for their inferences."

But of course, any published study, whatever the validity of its conclusions, represents a considerable outlay of money, time and thought. Therefore, it is necessary to consider what of value, if anything, can be retrieved from a published study when it is discovered to be flawed. That is, to consider whether the study can be salvaged by any post hoc maneuvers such as a different analysis of the data, editing the data, augmenting the data, or integrating the data in a meta-analysis.

Those studies in group B3) can often be salvaged by re-analysis of the data. Those studies in group B4) may only be salvaged by pooling the data from several, often similarly flawed, studies of a common dose-response relationship and then re-analyzing the pooled data. (Pooling the data is, of course, meta-analysis in its rawest form.) In the subsequent re-analysis it is important to be able to discriminate between random and systematic effects in the between-study differences should they prove to be significant. For example, the slope and/or intercept of a dose-response curve may prove to be study-specific as in the case of the Poisson models of the Tradescantia data. Therefore, the model of the pooled data must include indicator variables to identify the presence of a study-effect and its interaction with other covariates as described above.

For secondary analysis of a study to be possible and for the study weaknesses (of design and/or analysis) that it may disclose to be corrigible, either by salvage maneuvers such as data augmentation and mixed estimation, or by a larger meta-analysis, the investigator must have access to "accurate and circumstantial accounts" of the original primary observations in each of the published target studies. Thus, a key desideratum in the growth of knowledge is that the editors of scientific journals require that access, by the readers, to the primary data of a published study be a condition of acceptance of a report of the study in their respective journals. This requirement could be satisfied most expeditiously by simply including in the published paper the primary data, together with an adequate account of the circumstances in which they were obtained, that would permit any reader who wished to reproduce the results to do so. If the data set, etc., were too large for publication to be an economically feasible solution, then the same requirement could be met by a note in the published paper stating that the data, etc., are available upon request - and at a nominal cost - from the first author.

12. Are data taken seriously?

"The proper function of data is to change opinions. That is, to map prior opinions into posterior opinions."

E. Learner, 1970

"Models are not the source of information, however; data are ..."

R. Thisted, 1971

An evaluative study such as the present report must inevitably disclose not only what a scientific group believes but also what it values as an evidentiary basis for its beliefs. One finding from the reviews described in Annexes II-IV, which we briefly mentioned above, that seems to be quite remarkable, in a Baconian science such as radiation biology,[12] is that the primary data on which the findings of the published reports are based do not appear to be regarded by their authors as having much probative value. It appears that their data are not, that is, taken very seriously as empirical evidence for their findings by many investigators.

Since this finding seems to suggest that several centuries of scientific thought and (generally successful) scientific practice have been set aside, it seems necessary to recapitulate some of the empirical evidence for it:

a) Tables of the primary data on which the reported findings are based are rarely published in any report, including the reports of such groups as the NCRP and NAS/NRC. And the primary data of a report are usually not otherwise accessible to the average reader. (Requesting them from the first author is usually - though not always - futile. Or, when successful, may present the reader with an experience of the so-called Wolins Effect. (Wolins, 1962). For example, it required nearly two years for the principal author of the TG1 report to obtain the version of the LSS cancer mortality data that provided the parameter estimates in Table V-9 of the BEIR Report.) Instead, the report usually only provides graphical summaries of the data in which the primary observations are wrapped up in such combinations as NSD, TDF, D/N, % response, etc., that cannot be subsequently unpacked by any reader. It is the case, of course, that the simple "presence" of any "data" - in any form whatever - in a report tends to reinforce the prior beliefs of the reader concerning the matters addressed by the study, regardless of whether the data that are presented in the report confirm those beliefs, confute those beliefs, or are quite irrelevant to those beliefs. Thus, the data appear to serve an iconic rather than an evidentiary role - as we shall make clear in b)-l) below.

b) Statistically adequate evidence - such as provided by aggregate (e.g., Pearson χ^2) and case (e.g., residuals and other diagnostics) statistics of goodness-of-fit - that the model that is deployed is consistent with the data that are offered is rarely presented in a published report. Most often the question of goodness-of-fit, or concordance, of the chosen model and data of the study is simply ignored in the published reports. The authors evidently assume without checking - and require the readers to accept without recourse - that, for the purposes of the report, the chosen model provides an adequate "fit" to the data cited in the report. This weakness was remarked by Bacon some 350 years ago: "The human understanding from its peculiar nature, easily supposes a greater degree of order and equality in things than it really finds, and although many things in nature be sui generis and most irregular, will yet invent parallels and conjugates and relatives, where no such thing is ..." (Bacon, 1620).

c) The information on the inherent form of the distribution of the random part of the response (Normal, Poisson, Binomial, etc.) that is provided by the data is ignored by many investigators. For example, if the response described by the data is a proportion, say r_i responders in n_i at risk, then the distribution of the response must be Binomial. And the form of this distribution not only provides the inherent weight of the respective observations within the sample but often, as in the case of the Binomial response, provides the inherent form of the dose-response curve (e.g., sigmoid) - or surface - for the model as well. The appropriate model of Binomial dose-response data has a link function that is either probit, $z_i = \Phi^{-1}(\pi_i)$, or logit, $z_i = \log[\pi_i/(1-\pi_i)]$; where $0 \leq \pi_i \leq 1$ and r_i/n_i is the sample estimate of π_i. Too often, however, investigators assume that the link function is the identity, $z_i = \pi_i$.

d) There are relatively few primary studies that provide sets of data bearing on any given issue that have been found to be acceptable to the peer-group.[13] In every field these few sets of observations provide most of the empirical basis upon which some hundreds of studies had to be - and have been - published. Six of the more familiar of these "canonical" data sets are: i) the Cohen (1966) collation of the Ellis (1942), Jolles and Mitchell (1947), and Patterson (1959) isoeffect data on clinical radiation response of Normal tissues; ii) the Sparrow et al (1970) dose-response data on radiation mutagenesis in plants; iii) the Shellabarger et al (1969) dose-response data on radiation-induced mammary neoplasia in rats; iv) the van der Kogel (1979) dose-response data on radiation-induced hind leg paresis in rats; v) the von Essen (1963) isoeffect data on "volume effects" in humans; vi) the LSS (1980) dose-response data on radiation-induced leukemia and cancer in humans; vii) the ankylosing spondylitic (1980) dose-response data on radiation induced cancer in humans.[14]

However, it is most important to recognize that all of these data sets, the experimental as well as non-experimental, are flawed in consequential ways, that resulted in either severe weakening, or a complete refutation, of the received models that were - or could be - constructed on them. Thus, one should also bear in mind Fisher's remark: "If the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too."

e) In addition to the paucity of primary studies and the absence of accurate and circumstantial accounts of the primary data in the published report of the study, the presence of data-analytic solecisms, and the fact that important information on the radiation response of interest that is manifest in the data is manifestly ignored by the investigator, as noted above, perhaps the most persuasive evidence that their data are not taken too seriously by most investigators is the fact that most studies seem to be designed to obtain so little of it! Indeed, most studies may be said to be "data-starved". Thus, in the case of the randomized clinical trials (RCTs), the experimential studies which currently represent the most scientifically sophisticated level of clinical research, the lack of adequate numbers of patients in the respective treatment arms of the trial is repeatedly remarked in recent evaluations of these trials. For example, Klein et al (1986) note that, "The average sample size per treatment arm of these studies was 21 and only 2 of the 28 trials had more than 50 patients per treatment arm." But Zelen (1982) has previously recommended that, "Comparative trials should be planned with a minimum of 100-200 patients per treatment. Trials with fewer patients are likely to produce more false positive results than true positive results." Of course, with small sample sizes the false negative error rates, as well as the false positive error rates, increase dramatically: "... small studies are more prone to bias and the bias is substantial" (Berlin, Begg, Louis, 1989). "An implication of this study is that the results of small published studies are typically unreliable even taking into account the fact that such trials are imprecise due to sampling variation" (Berlin, Begg and Louis, 1989). And, Randomized controlled clinical trials are often too small to detect clinically important differences. (Sacks et al, 1987).

But too few data to resolve the issue are not only a weakness of the more expensive clinical trials but also of the less costly laboratory studies. For example, as nearly as we can determine, in all of the recent (– 1980) radiation toxicity studies in which the response has a Binomial distribution the numbers placed at risk, n_i at each treatment level $\underline{x}_i^T$ are too small: $n_i \simeq 5$ rather than $n_i = 25 - 30$ required by a statistically adequate design. (See, for a typical instance, the hind leg paresis data of Annex II, part 3.) It is of interest to note that the sample sizes in clinical trials and in laboratory dose-response studies at high doses are both smaller than recommended by a factor of about five. Curiously enough, in earlier (– 1960) dose-response studies of radiation carcinogenesis at low doses, in which the response also has a Binomial distribution, the numbers at risk are more nearly adequate: n_i – 25. (See, for example, the mammary neoplasia data in Fig. 13a and Annex III, part 6.) However, for both toxicity and carcinogenicity studies it is most often the case that the selected range of dose in the experiments is "covered" by an experimental design in which the levels of dose increase non-uniformly, in geometric rather than in arithmetic increments. The non-uniform design does, perhaps, provide the more economical coverage of the region of interest. The deleterious effects of non-uniformity in the distribution of dose levels may, in some studies, be exacerbated by non-uniform replication of observations at each level of dose so that there are more observations at some levels of dose than at others. See, for example, the mutagenicity data Fig. 1 and in Annex III, part 5. This results in non-uniform distributions of dose and observations that are characterized by the presence of extreme - and thus "influential" - observations in the data since in both types of study the received models specify dependence of the response on dose, D, rather than log dose, x_1 = logD. (If the dose increases geometrically, then the logarithms of the dose increase arithmetically - i.e., the levels of $\underline{x}_i^T$ are uniformly distributed.) Of course, both of these weaknesses of experimental design - inadequate numbers, n_i, at risk and inadequate distributions of levels of dose - may be due to exogenous pressures to obtain publishable results as economically as possible rather than as accurately as possible. (But see Zipf's principle below. Zipf, 1965.)

Economic principles and practices have, to be sure, enjoyed fat roles in the thought and practice of physical science. They provide the conceptual basis for the several teleologies such as Maupertuis' Principle of Least Action and Fermat's Principle. Ockham's Razor is another obvious example. And, somewhat less remotely, Mach, for instance, is often quoted as having said that the choice of a scientific law is based on 'economy of thought' or 'economy of description' of observations" (Jeffreys, 1960). However, several modern radiation scientists seem to have confused Mach's economy of description of observations with economy of production of observations since their experimental designs can be characterized as placing a minimal number of animals at risk at each of a minimal number of levels of treatment variables that can cover the selected ranges of predictor variables.

Given the profound effects of the size of the estimation sample on the cost of the study as well as on the quality of the admissible inferences which may be made from regression models thereof, it seems also appropriate to repeat here the celebrated animadversions - reported as, "The Law of Small Numbers" - by A. Tversky and D. Kahneman of the universities of Stanford and British Columbia, respectively, on the evidential value of small samples of observations. (Tversky and Kahneman, 1982a).

"The Law of Small Numbers

"The law of large numbers guarantees that very large samples will indeed be highly representative of the population from which they are drawn. ... People's intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well."

"... the believer in the law of small numbers practices science as follows: 1) He gambles his hypothesis on small samples without realizing that the odds against him are unreasonably high.

2) He has undue confidence in early trends (e.g., the data of the first few subjects) and in the stability of observed patterns (e.g., the number and identity of significant results. He over-estimates significance.

3) In evaluating replications, his or others', he has unreasonably high expectations about the replicability of significant results. He underestimates the breadth of confidence intervals.

4) He rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal 'explanation' for any discrepancy. Thus, he has little opportunity to recognize sampling variation in action. His belief in the law of small numbers, therefore, will forever remain

intact."

It is worth noting that these weaknesses in the received methods and measures of experimental designs - and also in the received concepts, methods, and criteria of data analysis as described above - are consistent with a general extremum principle of human behaviour, namely, Zipf's Principle of Least Effort which asserts that any person, "... will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems." (Zipf, 1965). To borrow a phrase from Fischoff et al (1982), "Good data analysis is so rare because poor analysis is so easy." Zipf's principle provides an explanation of the tendency of the investigator to obtain his results with the minimum of effort rather than the maximum of accuracy (vide supra) and also why the investigator will not commit himself to acquire the appropriate data analytic techniques but instead, per Kuhn, commits himself to lifelong, "strenuous and devoted attempts to force Nature into the boxes provided by his early professional education". It should be remarked that Zipf's Law (of the hyperbolic distribution of word frequencies) is regarded as better founded than is his Principle. Nonetheless, the latter seems to offer a coherent explanation of what otherwise often seem to be inexplicable behaviours.

f) The inherently multivariate nature of the matrix of covariates, X, is ignored. This omission leads the investigator to ignore the effects - inflation of $\hat{\beta}$ and of Var($\hat{\beta}$) - of the presence of high levels of multicollinearity in the sample distribution of predictor variables on the estimates and inferences on the parameter vector of dose-response models that are obtained from the sample. Moreover, ignoring the multivariate nature of the response leads not only to ignoring the effects of multicollinearity but also those of under-fitting as well. (See, for example, the studies in Annexes II and III.)

g) In three studies, one on volume effects, one on radiation carcinogenesis, and one on radiation toxicity, we have found that the updated version of Mach's principle of "economy of thought" - parsimony in observation - has been carried to a logical extreme: Some of the data from which the respective models of radiation response that are described in these reports were constructed are dispensed with altogether - a maneuver that achieves economy of thought since there are fewer observations to think about. In two of these studies we found that fictive observations - "non-experiential data" - provide most of the empirical bases of the estimates and inferences on the respective dose-response models that are reported in the two studies and are repeated in subsequent authoritative reviews. These two studies are examined in Appendix I and in Annex II, part 3, respectively, to this report. In the former study a regression model was constructed of a data set that included hypothetical observations. In the study analyzed in the latter both the numbers at risk, n_i at x_i^T, and the numbers of treatment levels, that were presented in the report of the study that was originally published were found to be fictitious. In a third study, two (of seven) original observations - those at the lowest and highest doses - were deleted in order to force the received model on the data, a maneuver that forcefully recalls Acton's remark: "Some data fit lines naturally, but others have lines thrust upon them." (Acton, 1959)

h) The evidence of the sample data on the appropriate forms of the deterministic and random parts of the response is ignored, leading to misspecification of either the structure of the deterministic part, μ_i, or of the form of the distribution of the random part, ϵ_i, of the observed response, $y_i = \mu_i + \epsilon_i$, $1 \leq i \leq n$. Or both. Thus, we have found instances in the literature in which the model was overfit leading to inflation of the variance-covariance matrix Var($\hat{\beta}$) of the parameter estimates $\hat{\beta}$ of the model and of the bias and variance of estimates, f($\hat{\beta}$), of nonlinear functions f(β) of the parameter vector of the model (See the evaluations of the L-L and LQ-L models of leukemia incidence in Annex III. The BEIR III (NAS/NRC, 1980) data show a clear, preference for the L-L model on the evidence of both aggregate and case statistics of concordance, but the LQ-L is the "model of choice" in the BEIR III report). We have also found instances in the literature in which the model was underfit, leading to biased (aliased) estimates of β and deflated estimates of Var($\hat{\beta}$), i.e., the precision of estimate for some parameters, $\hat{\beta}_j$, is larger than would be the case if the model were more consistent with the data (See the evaluations of the single and multifraction LQ models in Annex II, parts 5 and 3. In the former a term in D^3 is omitted; in the latter a time factor is omitted).

i) In the received praxis, the sample data on radiation response are regularly transformed, transmogrified, down-weighted, (including being assigned zero weight), etc., in order that the observations may be "fit" to an equation of the form $y = a + bx$ in which the estimates of the parameters can be obtained, either graphically ("by eye"), or by ordinary least squares methods. This practice has let to some rather bizarre results in the construction of LQ models of cell survival: The evidence of the sample data is discounted by a) the imposition of ad hoc a priori information $e^{\beta}0 = m_1$, where m_1 is the survival at zero dose and $\psi = [0]$ on the parameter vector of the LQ model (Fertil and Malaise, 1981); b) the imposition of ad hoc weights, $w_1 = 10^6$, $w_i = D_i^{-2}$, $2 \leq i \leq n$ (Chapman, 1980) on the observations. The respective variance-covariance matrices of the parameter estimates of the LQ model of cell survival obtained by Poisson regression (Frome, 1968), least squares regression (Fertil and Malaise, 1981) and the Chapman plot (Chapman, 1975) are $(X^TW_FX)^{-1}$, $[\phi X^TX + R^T\psi R]^{-1}$ and $(X^TW_CX)^{-1}$, where $W_F = \mathrm{Diag}[m_ic_i]$, $\psi = 10^{-6}[I]$, $R = (1,0,0)$ and $W_C = \mathrm{Diag}[D_i^{-2}]$. Obviously, only the Frome estimate is correct for the Poisson data. Since they include the fictitious dispersion matrices ψ and W_C, the Fertil and Chapman covariance matrices are fictitious and thus confer specious levels of precision on the respective parameter estimates. Moreover, the imposition of the specific prior information on $\underline{\beta}$ that is entailed in the Fertil and Chapman methodologies also degrades the fit of the model to the sample since $RSS^{**} = RSS + (\hat{\underline{\beta}} - \bar{\underline{\beta}}^{**})^TX^TX(\hat{\underline{\beta}} - \bar{\underline{\beta}}^{**})$ in the Normal theory example, that is, their practice arbitrarily discounts the sample information on the model at issue.

j) The range of the predictor variables (e.g., D or D/N) in the study data and the range of these variables over which the model is stipulated to be valid often do not "intersect". For example, in one study on radiation toxicity 9 of the 19 observations lie beyond the stipulated range of validity of the multifraction LQ model of radiation toxicity: $10 \leq D/N \leq 20$ vs $0 \leq D/N \leq 10$. For another example, in one study on radiation mutagenesis the single fraction LQ model is stipulated to be valid over the range $0 \leq D \leq 100$ rads, however, the data include only 3 (of 20) observations above 24 rads, or 85% of the observations are included in the lowest 25% of the stipulated range of validity of the model. See Figs. 9 and 1a, respectively.

k) There are a wide variety of still other idiosyncratic and pathological ad hockeries to be remarked in the received treatment of the data in radiobiology modelling. For example, while in the received practice for construction of LQ models of cell survival (Fertil and Malaise, 1981) the observation at D=0 is assigned an infinite weight, $w_1 \longrightarrow \infty$, in the received practice for construction of linear models of the mammary neoplasia data (Shellabarger, 1969) we found that this observation was assigned zero weight, $w_1 = 0$. See Figs. 12b and 13a, respectively (see also Annexes IV, part 5 and III, part 6, respectively). Similarly, while the (received) methodology, described by Fertil and Malaise (1981), deployed in the construction of LQ models of cell-survival forces the dose-response curve to coincide with the response m_1 at the lowest level of predictor variable, $D_1 = 0$, (see Fig. 12b and Annex IV, part 5) the choice of experimental design deployed in the construction of LQ models of radiation toxicity forces the isoeffect curve to coincide with the response at the highest level of the predictor variables, $D/N \simeq 15$ Gy (see Figs. 22 and 23 and Annex III, part 3). And, of course, as we have remarked above, much of the information that is contained in dose-response data such as represented in Fig. 9 is discarded by isoeffect models of that data such as represented by the F_e-plot of Fig. 10.

l) In many studies a priori information on either or both the deterministic and random parts of the observed response that is either invalid, irrelevant, (or both), or inconsistent with the sample data, is introduced into the analysis. A vivid example of thoroughly implausible prior information on the deterministic part that has been deployed in a published study may be found in Tucker and Thames, 1983 (See Fig. 11c and Annex IV, part 4) in which the slope of a sigmoid dose-response curve for a binary (Binomial) response is constrained to be equal to zero at $\pi = 0.50$ for dose $D < 0$. (See Fig. 10c above.) But clearly $D \geq 0$ and the slope is a maximum at $\pi = 0.50$ (or, about as far from zero as it is ever going to get). An equally vivid example of the deployment of prior information on the parameter vector $\underline{\beta}$ that is inconsistent with that of the sample is found in Supe et al (1983) in which, in order to obtain an isoeffect model for extrapolation, $x_1(\pi) = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$, the a priori constraint $\alpha_1 + \alpha_2 = 0.24$ is imposed. However, we have shown both by graphical display and by the γ-test of the Mixed estimation procedure (See section 7.2.4) that this constraint is inconsistent with the information on $\underline{\beta}$ contained in the sample data. See Fig. 36a and Annex IV, part 3.

Usually, in the case of models of the deterministic part, dubious prior information is introduced by the investigator as a linear constraint, $\underline{r} = R\underline{\beta} + \underline{v}$, $Var(\underline{v}) = \psi = [0]$, on the parameter vector, $\underline{\beta}$. The investigator assigns sufficient weight to this non-sample information, that is, $\psi \equiv [0]$ a priori, so that it tends to dominate the cognate sample information despite both its perhaps questionable relevance and its questionable ontological status. It is important to note that this description refers to the effective behaviour of the investigator(s), that is, the received data analytic practices are equivalent to Mixed estimation with the constraint $\underline{r} = R\underline{\beta} + \underline{v}$, $Var(\underline{v}) = \psi = [0]$, in that the two practices achieve the same estimates. However, there is no evidence that these investigators were at all aware of the real nature or meaning of their practice. Kuhn's 1977a observations on the practices of physical science are quite appropriate to the current practice of radiation biology: "... in scientific practice as seen through the journal literature, the scientist often seems to be struggling with facts, trying to force them into conformity with a theory he does not doubt." The results of those struggles that are reported in the literature we reviewed, show theory to be a clear winner.

Thus, it should come as no surprise to find that for each published study that we analyzed by normative statistical concepts, methods, and criteria, in which, in Ehrenberg's locution, the study data are permitted "to speak" to the matter at issue, (Ehrenberg, 1975) and in which the prior information on the response metric is both correct and correctly represented and implemented, the received model of radiation response failed to be as consistent either with study data, or with generally accepted a priori information on the response that was the object of the investigation, as did the cognate rival model that we introduced for comparison - and for control. For example, the respective linear-quadratic (LQ) models of radiation leukemogenesis in humans (LSS sample, T65D dosimetry), hind-leg paresis in rats, bone marrow stem cell survival in rodents, radiation mutagenesis in Tradescantia, and the linear model of radiation induced mammary neoplasia in rats were found to be either rejected by the data of the respective study, on the normative statistical criteria of concordance, or were found to be much less concordant than was the rival model - again on the evidence and criteria of normative statistical methods: i) Analyses of residuals (including the sampling distribution of a residual sum of squares). Compare the respective chi-squared residual plots for the rival models M1a, M1b and M2 of radiation toxicity data and for the rival models T and LQ of cell survival data in Figs. 16, 17, 18, 20 and 26 of the present report and Annex II, parts 3 and 5; ii) Tests of invariance of the form and/or parameters of the model between the data of different studies; Compare the respective degrees of invariance of the T and LQ models and parameter estimates of cell survival in Annex II, part 5. iii) Consistency of the shape of the dose-response curve with generally accepted a priori information. Compare the respective curvatures, $k = (d^2y/dx^2)/[1+(dy/dx)^2]^{3/2}$, in the region $0 \le D \le 2.5$ Gy for rival models T ($k < 0$) and LQ ($k > 0$) of the cell survival data in Figs. 27a and 27b. A priori evidence suggests that $k < 0$ in this region, signifying the presence of a "shoulder" (and hence of either redundant cellular architecture or repair processes) is correct. This finding suggests that the LQ model is inappropriate. iv) Evaluation of scope. For example, in Annex III, part 6, for the mammary neoplasia data, the rival probit model has greater scope than the received linear model since it accounts for observations that are not accounted for by the linear model. In the case of the latter received model, the observations at the extremes of dose $D = 0$, $D = 500R$ were each assigned zero weight - deleted - in order for the linear model to fit the sample data in the original report. See Annex III, part 6. We found the dose-response curve for mammary neoplasia to be bi-phasic rather than monophasic (two regions of the data with different slopes rather than a single slope). Moreover, these latter analyses based on the rival probit model disclosed strong empirical evidence against the received non-threshold linear hypothesis of cancer induction by low

LET radiation and against the received hypothesis that the neutron RBE is dependent on dose. These two latter beliefs, of course, guide much of the current regulatory policy in this field. But, "The existence of carcinogenic thresholds can, in general, be neither proved nor disproved by the conventional bioassay; however, the concept of a threshold has worked well in the past for many toxicological responses. The E-NOEL [effective no observed effect limit] and its uncertainty should be reported along with other estimates of potential risk when the model is shown to give an adequate fit to the data. Risk managers should be made aware when the existence of a threshold is consistent with the data." Of course, "... there is a tension between honesty and prudence .. Probabilistic reports about adverse consequences to health are very often slanted to be conservative. I am arguing that it is better to report honestly, and that prudence should appropriately be represented in the evaluation process - not in the assessment process." (H. Raiffa, 1982). The linear model of mammary neoplasia is clearly the more prudent model; however, in the present context it provides an example of a motivational error. (And evidence of a belief in final causes, the principle of least risk as well: The "true" model is that which minimizes risk.)

A careful review of the data suggests that these latter two beliefs - non-threshold linear hypothesis and dose-dependence of neutron RBE - may be instances of what we have referred to as "methods instigated" hypothesis in Annex II, part 5. The re-analyses of received data on these issues that were presented in Annex III, part 6 disclose that a probit model, $z = \beta_0 + \beta_1 x_1$, where $x_1 = \log D$, provided a statistically adequate fit to both the gamma and neutron-induced responses, with $\beta_1(\gamma) = \beta_1(n)$, suggesting that the RBE is independent of dose. The extended probit model $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where $x_2 = 0$ (neutron), $x_2 = 1$ (gamma) provided a statistically adequate fit to the pooled $(\gamma + n)$ data and enabled us to obtain estimates of both the bias and variance in the sample estimate of $\theta = \beta_2/\beta_1$, the so-called neutron RBE. See Fig. 32.

The findings of such anomalies as described above suggest that the current paradigm under which research in radiation biology is informed and conducted - that is, the theory, methods and criteria deployed by the group in the practices of construction, discrimination, and validation of dose-response models - has been greatly weakened by the ineluctable Law of the Instrument which states that both the identification of, and solutions to, the problems encountered in any enterprise are limited by the tools at one's disposal (Or, as Maslow has remarked, "If the only tool you have is a hammer, you tend to treat everything as if it were a nail."). Apparently the only tool available to many of the investigators whose reports appear in Table 1 is one that permits one to "fit" the model $y = a + bx$ to data.

It has long appeared to us that some "re-tooling" of the data acquisition and analysis parts of the modelling enterprise in radiation biology would be timely, fruitful and cost-effective. In the Kuhnian model of the scientific enterprise the occurrence of such paradigm failures as we have described in this report signal the presence of crisis in the practices of the group that can only be resolved by adoption of a new paradigm by the group: "As in manufacture so in science - retooling is an extravagance to be reserved for the occasion that demands it. The significance of crises is the indication they provide that an occasion for retooling has arrived." (Kuhn, 1970a). Or, "The function of crises in the sciences is to signal the need for innovation, to direct the attention of scientists toward the area from which fruitful innovation may arise, and to evoke clues to the nature of that innovation." (T. Kuhn, 1977). The suggestion of a crisis in radiation biology disclosed by the secondary analyses summarized above and described at length in Annexes I to IV, suggests that the "re-tooling", or "innovation" required by Kuhn's model of how science works, could be achieved by the adoption by the radiobiology peer group of the concepts, methods, and criteria of statistical modelling in the acquisition and primary analysis of their data.

13. Criticism and quality assurance. A new relationship of science and the law.

"... we can learn from our mistakes."

"The way in which knowledge progresses, ... especially ... scientific knowledge, is by unjustified (and unjustifiable) anticipations, by guesses, by tentative solutions to our problems, by conjectures. These conjectures are controlled by criticism; that is, by attempted refutations, which include severe critical tests."

K. Popper, 1965a

13.1 "Criticism and the Growth of Knowledge".

"The growth of scientific knowledge, then, does not depend so much on the honesty as on the skepticism of scientists."

W. Schmaus, 1987

The results of the secondary analyses described in Appendix I and Annexes II-IV strongly suggest that those received beliefs of radiation biology that are expressed in those passages that were quoted in part 2 ("What do we believe?") of this report are not at all well-founded. At the very least it must be clear that the debate on many of these issues - concepts, methods, criteria, and models - has been prematurely foreclosed. However, such "negative" results cannot be too surprising in the context of other assessments of the published biomedical literature such as that by Williamson et al (1986): "The findings reported in the 28 assessment articles suggest that the average practitioner will find relatively few journals articles that are scientifically sound in terms of reporting usable data and providing even moderately strong support for their inferences."

Nonetheless, it may be reassuring to recall that the negation of the positive findings of the studies listed in Table 1 is, logically, a more "positive finding" than their confirmation would be: "Positive and negative results in experiments are not equivalent in their logical implications. In fact, while they have unquestionable bearing on the subjective aspects of belief, successful experiments have no logical bearing on the truth status of their source (i.e., a theory or hypothesis). As counter-intuitive as this may seem, it is a clear consequence of logical analysis ... It is only negative results (contrary-to-prediction) in experiments which carry logical implications" (Mahoney, 1977).

These logical implications, remarked earlier by Bacon in the passages quoted above in part 5, have been variously described more recently as follows:

a) "... the asymmetry of a generalization and its negation in their relation to empirical evidence. A scientific theory cannot be shown to apply successfully to all its possible instances, but it can be shown to be unsuccessful in particular applications" (Kuhn, 1970b).

b) "... the asymmetry between verifiability and falsifiability; an asymmetry which results from the logical form of universal statements. For those are never derivable from singular statements, but can be contradicted by singular statements. Consequently, it is possible by means of purely deductive inferences (with the help of the modus tollens of classical logic) to argue from the truth of singular statements to the falsity of universal statements. Such an argument to the falsity of universal statements is the only strictly deductive kind of inference that proceeds, as it were, in the 'inductive direction', that is, from singular to universal statements" (Popper, 1965a).

c) "Only the falsity of a theory can be inferred from empirical evidence, and this inference is purely a deductive one" (Popper, 1965a).

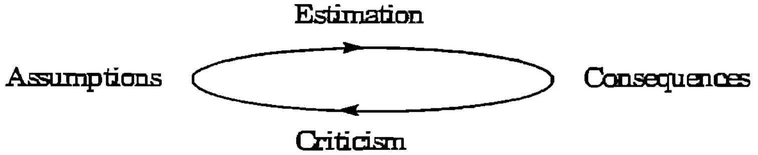
But, of course, in the real world, as Popper (1965b) remarks, "In point of fact, no conclusive disproof of a theory can even be produced; for it is always possible to say that the experimental results are not reliable, or that the discrepancies which are asserted to exist between the experimental results and the theory are only apparent and that they will disappear with the advance of our understanding." This is echoed by Kuhn (1970b): "... where a whole theory or often even a scientific law is at stake, arguments are seldom so apodictic. All experiments can be challenged, either as to their relevance or their accuracy. All theories can be modified by a variety of ad hoc adjustments without ceasing to be in their main lines, the same theories. it is important, furthermore, that this should be so, for it is often by challenging observations or adjusting theories that scientific knowledge grows."

However, Popper (1965b) notes that, "If you insist on strict proof (or) strict disproof) in the empirical sciences, you will never benefit from experience, and never learn from it how wrong you are." This is, of course, the crucial lesson; as Bacon (1620) has remarked, "Truth emerges more readily from error than from confusion." (In the construction of dose-response models the regression diagnostics clearly identify and measure "error" and thereby help "truth" to emerge from error.)

As have many others have, both before and since, Popper (1965b) has proposed that empirical science must be uniquely characterized by its process or method. The process that Popper proposes is that of conjecture and refutation: "... an alternating series of speculative conjectures and empirical refutations" (Popper, 1965a). Popper (1965a) bases the whole of his Logic of Scientific Inference on the disarmingly simple premise "... that we can learn from our mistakes." (As Kuhn (1970b) subsequently notes, the problem lies in the identification of a mistake.)

G.E.P. Box has prescribed the same method for the construction of useful models ("All models are wrong but some are useful."). Box (1984) has remarked that there are two kinds of inference required in the construction of useful models: "One kind of inference that may be called criticism involves the contrasting of what might be expected if the assumptions A of some tentative model of interest were true with the data yd that actually occur. This is conveniently symbolized by subtraction: y_d - A. The other kind of inference, which may be called estimation, involves the combination of observed data yd with the assumptions A of some model tentatively assumed to be true. This process is conveniently symbolized by addition: y_d + A." Box further points out that the methods, e.g., goodness-of-fit, regression diagnostics, etc., for model criticism can be best motivated and justified by Sampling Theory whereas, the methods, e.g., least squares, maximum likelihood, mixed, robust and ridge estimation, for model estimation are better motivated and justified by Bayes' theorem.

Initially Box (1979) and more recently Cook and Weisberg (1982) and McCullagh and Nelder (1989) have described the process of construction of (useful) linear models of sample data (obtained in either an experiment or a non-experiment) as proceeding by iteration of the basic cycle:



The sequence of iterations is terminated when a model has been constructed such that the sample of data is mapped into a "white noise" sequence by the model (Box, Hunter, and Hunter, 1979). Such a model will have retrieved all of the "signal", or information, that lies within the data.

Both philosophers of science, e.g., Bacon, Kuhn and Popper, and practicing statisticians, e.g., Box and Pearson, have described the fundamental role in the growth of knowledge that is played by informed criticism (to be distinguished from unreflective doubt): If we are to learn from our mistakes (pace Popper, 1966a) then there must be methods and criteria by which "mistakes" may be identified and their etiolgy and effects disclosed and corrected. In the case of dose-response models these methods and criteria are, as Box has pointed out, the methods and criteria of statistical modelling that we have described in this report; For example, the statistics χ^2 and D, sums of squared chi-square or deviance residuals, respectively, that measure the goodness-of-fit of model and sample; the statistic γ, that measures the concordance of sample and a priori information on $\underline{\beta}$ or $\underline{x}^T\underline{\beta}$); the regression diagnostics, etc.[15]

But in this respect we have described the total absence, in several well-known and frequently cited studies, of any evidence that suggests that the normative critical judgments were exercised in the construction and deployment of the received models. (It may be said that, in these cases, disbelief was not only suspended, it was hung by the neck.) Thus, a reader cannot be certain that it is scientific knowledge that is being reported in many of those papers that we reviewed - since there is no evidence of the presence of the process that leads to scientific knowledge. It may be appropriate to recall, by way of homely analogy, that the only warrant for the random nature of a "random sample" - and hence for the validity of the estimates and inferences subsequently derived therefrom - resides in the non-sample, or a priori, evidence that the process that generated the sample was itself a random process. So it is with the results of a scientific study; the only warrant for their believability is the evidence that they were obtained by a scientific process.

13.2 Science and the law.

There is still another aspect of the issue of the quality of the intellectual product that is retailed in much of the scientific literature, and, most especially, of the process by which that literature is produced, that should be called to the attention of the readers of this report. Most readers are, of course, by now quite familiar with the current pre-occupations, in several judicial, as well as academic and legislative bodies, with the provenance and probative value of the scientific literature that were referred to in the Introduction and summary of this report, i.e., the several causes celebres associated with the names of Slutsky, Stewart, Baltimore, Dingle, et al. However, many readers may not yet be quite so familiar with another more insistent, fundamental, and pervasive scientific concern of the judiciary that arises in the currently changing relations of science and the law. This concern is with the scientific process and its products and their respective evidentiary roles.

Currently, it appears that science and the law intersect most often in the testimony of the expert witness. Here it should be recalled that the role of the lay witness is to inform the court of the facts in the case as he actually perceived them. However, his (her) opinions as to the interpretation of these facts are inadmissible as evidence by law. On the other hand, the expert witness is required to deliver his (her) opinion as to how the facts in the case are to be interpreted. In particular, the expert witness may be expected to express opinions on the issues of causality. An expert testifies on his (her) conclusions, which are based on those facts and principles that are held to be generally accepted in his field, and on the facts in evidence in the matter at litigation. The experts' line of reasoning connects these latter facts to the conclusions expressed in the opinion that represents his testimony. The court must rule on the admissibility of the opinion of the expert witness on a case by case basis. This requirement has long burdened the "triers of fact" with, "... the need to evaluate expertise while simultaneously depending on it." ... "Judges, however, have generally refused to probe the validity of the reasoning behind a scientist's conclusions or to hold experts to the standards and criteria of scientific practice." (Black, 1988a). Instead, courts have relied on surrogate factors to validate the experts' expressed opinion that do not always guarantee scientifically valid evidence. The more common are: a) the professional credentials of the expert; b) the certainty in his testimony that is expressed by the expert; c) the general acceptance of the principles upon which his testimony is based. (Black 1988a, b).

It is now widely acknowledged that the use of expert witnesses in the traditional adversary system of American jurisprudence has had increasingly disappointing results. This unhappy circumstance has lead officers of the court to remark disparagingly of liars, damned liars, and expert witnesses, in increasing order of unreliability. The reasons for such remarks are not hard to find. They were, in fact, accurately anticipated by Descartes about three hundred years ago (and by Cicero, nearly two millennia earlier). Three more contemporary articulations of these reasons are: 1) "The scientific community is large and heterogeneous, and a Ph.D. can be found to swear to almost any 'expert' proposition no matter how false or foolish"; 2) "An expert can be found to testify to the truth of almost any factual theory, no matter how frivolous, thus validating the case sufficiently to avoid summary judgment and force the matter to trial ... Juries and judges can be, and sometimes are, mislead by the expert for hire." 3) "Today practicing lawyers can locate quickly and easily an expert witness to advocate nearly anything the lawyers desire." (Black, 1988b; Graham, 1986).

In an attempt to reach verdicts in cases involving forensic science and medical testimony, e.g., the so-called toxic tort cases,[16] that are consistent with scientific reality courts now appear to be rather rapidly abandoning the adversarial system that is based on the conflicting testimony of opposing expert witnesses for a kind of inquisitorial system of active and critical judicial review of the scientific evidence on the matter at issue that is presented by the opposing counsels. In such reviews, the testimony offered in evidence is required to be consistent with, or, "... to conform to the standards and criteria to which scientists themselves adhere." Thus, "... the problems science poses for the law parallel and reflect the philosophical problem of defining science." [This is, of course, precisely the problem raised by the findings of the secondary analyses of those studies that are described in the present report.] "When a dispute about the admissibility of scientific evidence hinges on the validity of an expert's reasoning, an acceptance test based on the current, realistic view of science provides the only way to reach a rational and consistent decision. A court should determine which scientific fields are relevant to the dispute and then turn to the peer-reviewed literature from those fields for guidance. Experts should be required to make their reasoning clear and explicit and to explain how their conclusions derive from accepted scientific practice. An expert who cannot demonstrate that his or her reasoning conforms to the standards of science should not be allowed to testify." (Black, 1988a. Emphasis added) For scientific evidence that does not conform to accepted scientific practice, a court should either exclude the evidence as inadmissible or find it insufficient to sustain a verdict. However, as has been shown repeatedly in this evaluation, as well as in the evaluations of others that have been cited, the reports published in the peer-reviewed literature itself do not in every instance - some would say rarely - conform to accepted scientific practice.

Such a judicial review raises two of the same questions of the validity of the principles, methods, and criteria deployed in the practices of a professional peer-group that motivated the work of the Task Group 1 as is described in this report: 1) Why do we believe it? 2) Should we believe it? If the courts seek to resolve evidentiary disputes about science by reference to the scientific literature (As indeed they already have in some cases: A court recently rejected scientific evidence because it had never been refereed or published.) this imposes the judicial requirement, "... to review the literature critically with an eye toward why articles are published and in what journals." (Black, 1988a). It is here that the critical distinction, drawn by Fleck, between exoteric and esoteric science, may cause the most difficulty for the scientists - as well as for the courts. It is clear that the weaknesses in radiobiological praxis that have been disclosed by the present review, as well as such reviews by others, raise the distinct, if disturbing, possibility that there will be found, on occasions of critical judicial review of that literature, papers which present empirical evidence that has previously supported a grant but cannot now sustain a verdict. It seems likely that occasions for critical judicial review may arise more frequently in the future than in the past. For, as several observers have remarked, radiation is a strong "litigen".

14. "What shall we do now?"

"When a major error is suspected in a particular scientific method, the customary scientific response is to instigate research that will allow the suspicion to be either confirmed or refuted."

A. Feinstein and R. Horowitz, 1982

14.1 Classical meta-analysis. "Validated reviews" of the literature.

"... to keep up with a given field, seek reviews based on validated research sources ... beware of any review that merely summarizes and does not validate its source information."

Williamson, et al, 1986

"All positive results should be independently confirmed ... Physicians in practice should exercise caution in adopting a new therapy if there is no independent confirmation."

M. Zelen, 1982

"Several decades ago, much of our work as scientists consisted of reproducing the research of other investigators and then going further." ... "Much of this has apparently changed; reproducing the experiments of other investigators is no longer of primary concern." ... "Who has the time, interest, money or need to reproduce another scientist's results?" ... "The implications of not reproducing experiments are severe. Much of what is published goes unchallenged, may be untrue, and probably nobody knows. Does anybody care? Do the methods used to obtain results matter any more? The foundation on which we based our research was other scientists' methods and results.

Now the foundation is trust. Are you comfortable?"

A. Neufeld, 1986

The answer to the initial question that this task group was formed to address concerning the nature and size of the losses to be incurred by deployment of the wrong model of radiation dose-response cannot yet be answered in the fullness that is required. However, our preliminary work strongly suggests that the question itself is a non-trivial one and that the losses are likely to be large. And, as remarked above, the comparison of the respective PRESS statistics of two or more rival models, gives a partial answer - or the "appearance of an answer", in Tukey's locution - on a squared loss criterion.Such comparisons disclose that for those data sets that we examined, the losses for the received models are not least - in that the losses for the rival model were found to be still less in each case. In the contexts of sections 10 and 13 of this report some may find this conclusion disturbing.

The secondary analyses described in Annexes I-IV and summarized above have disclosed the presence of debilitating weaknesses in the current radiobiological praxis, or in Kuhn's metaphor, that this praxis "no longer defines a playable game" (one must also ask whether it ever did). What should be an appropriate response to this disclosure? Perhaps the first immediate response should be to repeat our earlier disclaimer, namely, that the findings of this report must not be over-read to imply that none of the received models can be "believed" under any circumstances. Rather, our re-analyses disclose that the empirical evidence for these models that was presented in those studies which we examined is not beyond dispute by reasonable men and women. (Or at least not beyond dispute by those who, if not altogether reasonable, are reasonably well-informed and are endowed with reasonable levels of intellectual skill.) As we remarked earlier, it may be the case that other studies which we have not yet examined do provide incontestable evidence for the validity of the model(s) deployed therein - one recalls that, "absence of evidence is not necessarily evidence of absence".

But the role of the TG1 has been to determine the degree to which, on the basis of the data presented in a given study, the model proposed therein was defensible against rival models of that data. The received model must provide a defensible explanation, not merely a plausible one, since, "... there are different thresholds for the ascription of plausibility."

The TG1 has read its findings as suggesting the presence of a "Kuhnian crisis" in radiation biology. Probably the first step in the "re-tooling" that Kuhn has described as the normative response to the occurrence of crisis is to learn to take the data more seriously - but yet not too seriously (Leamer (1978) recommends the Bayesian view: "Let the data incrementally affect your opinion.") than presently often appears to be the case (It has been said that there are two kinds of scientists: Those who know too little about their data - and those who know too much). For example, one should never arbitrarily down-weight valid observations, nor fail to exploit relevant ones; e.g., in one oft-cited study on rat mammary neoplasia, two (of seven) observations were discarded, and in one oft-cited study on rat hind-leg paresis the data on dose response that was obtained in the study was ignored and an isoeffect model was constructed instead to convey the information on the dose-response relationship. Taking data more seriously requires, at the very least, 1) that one should learn to take more data on the matters at issue (e.g., in dose-response studies, place more than four or five animals at risk at each of more than two or three levels of dose - per dose-response curve; in isoeffect studies use more than two points to fit a straight line); 2) that one should learn to take more data more often (e.g., data "age",and the substantive conclusions of too many current publications are based upon weak data obtained 10-20 - even 30 - years earlier); 3) that one should learn to take more informative data on these matters (e.g., take more data that is less encumbered by the presence of outlying and influential observations, highly correlated variables, etc. do more dose-response studies - which examine several levels of (conditional) response - rather than isoeffect studies - which examines only a single level of response - since the elucidation of dose-response relations would seem to be the real "object of the exercise"); 4) that one should learn to publish one's primary data (primary data are now rarely published although they provide one of the stronger warrants for the validity of any finding: "The meaning of a data set is that it changes opinions. It takes particular prior opinions into particular posterior opinions. A data set may thus be fully described in terms of the mapping that it implies from prior distributions into posterior distributions." (Leamer, 1978).); and 5) that one should learn to take still more information from that data which one has already taken (e.g., "A regression [model] is constructed using prior knowledge, data, models, and a fitting (estimation) process of some form. It is important to know when the resultant regression [model] depends heavily on a small part of the prior knowledge, on a small part of the data, or on the exact choice of model or fitting process." (R. Welsch, 1986).

However, the required "retooling" is not likely to take place very soon. There are several reasons for this. One is that the awareness of either the presence or the nature of the apparent "crisis" has not penetrated very far into that peer-group; to date it seems to be confined to statisticians and to a few other scientists with an interest in statistical methods - as is consistent with the Kuhnian scenario: "Normal Science does not search for anomalies and, when successful, finds none" (Kuhn, 1962). Or, "it is impossible to impress ... people with truths they aren't ready to hear, much less accept." (L. M. McMurty, 1990).

Another reason is a behavioral epiphenomenon, most recently known by the mild pejorative "belief perseverance". See Annex I. The phenomenon was most eloquently described by the writer Tolstoy (1898): "I know that most men -not only those considered clever, but even those who are clever and capable of understanding the most difficult scientific, mathematical, or philosophic problems - can seldom discern even the simplest and most obvious truth if it be such as obliges them to admit the falsity of conclusions they have formed, perhaps with much difficulty - conclusions of which they are proud, which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives." Kuhn (1970a) has described its effects in the words of the physicist Planck: "... a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." Or, as W. Trotter has remarked more recently, "... a new idea is the most quickly acting antigen known to science. If we watch ourselves honestly we shall often find that we have begun to argue against a new idea even before it is completely stated." (And there appears to be no way to reverse the onset of intellectual anaphylaxis.) Thus, the required retooling may take at least a generation, perhaps two.

In the meantime there are at least three distinct, but complementary programmatic responses to our fourth question that should prove to be quite fruitful. All three are meta-analytic and thus depend heavily upon secondary analyses of published data - just the sort of work that is described in Annexes I-IV of this report. For, as Zelen (1982) has recommended, with respect to clinical trials, any innovation, including - or perhaps especially - any new model of radiation dose-response, should be independently verified before it is adopted for routine deployment in the clinic. See for example the reviews of the first misonidazole clinical trial in radiotherapy (Herbert, 1982); the BEIR III risk estimates (Herbert, 1986b); the MRI malignancy index (Herbert, 1986d, 1986e); the LQ model (Herbert, 1987). These latter represent earlier and abbreviated versions of the "validated reviews" of the literature that have been recommended by Williamson, et al (1986). Thus, the first response of the Task Group to the insistent question of "what to do" is to continue to provide both secondary analyses and - where possible - meta-analyses, that is, integrations, within both animal and clinical studies on dose-response models and across the entire spectrum of response - as in the present report. Therefore, we initiated secondary analyses of the set of papers most often cited as providing the original empirical evidence for introducing a time-factor in the linear form γT into the LQ model: 1) Travis and Tucker (1987); 2) Mah et al (1987); 3) Wara et al (1973).

As we remarked above in section 1.2, there is ample precedent for preparing such validated reviews in at least two other areas of applied science: a) Nuclear Physics (the so-called Particle Data Group) and b) Thermal Physics (the Thermophysical Properties Research Center). Both were founded about 1957. Such validated reviews would identify those papers which are most important to the group, and, as Derek Price has suggested, only half-facetiously, those papers which can be identified as "important" should be *republished* in a *Journal of Really Important Papers* with wide circulation and high profile so that they do not escape the attention of those who would most profit from reading them. Presumably those identified as *unimportant* can be regarded as being *depublished*? - or else remanded to the *Journal of Irreproducible Results*?

A bit less flippantly, the results of these secondary analyses should be published regularly in what may be usefully termed as "BEIR" reports - Bulletin of Evaluated Information in Radiobiology. These bulletins could be disseminated through an existing distribution mechanism such as the AAPM Report Series. There is an obvious cost-benefit aspect to this service. The mathematician, David Hilbert, has observed that the importance of a scientific paper can be judged by the number of previous publications it makes unnecessary to read. The Hilbert criterion has become more relevant in the face of the relentless production of research papers. (There are approximately 23,000 different biomedical journals currently published in the English language. I. Olkun, private communication.) Libraries must become more selective in what they retain - and retrieve. And readers must become more selective in what they read - and recall. More stringent criteria in quality assurance for modelling are required for both tasks.

Hilbert's remark also anticipates one of the non-epistemological rewards of meta-analysis, namely the relative cost-effectiveness of such studies themselves: In competing for study funds one may well ask the granting agency which is the more efficient use of (limited) resources: generating still more data in an (n+1)th experimental study of a given issue in radiation response, or a study that further analyzes, *evaluates* - and perhaps *integrates* where feasible and rewarding - the extant data and/or findings of the previous n studies?

These secondary analyses provide the basis for meta-analysis which may be considered to be a type of survey research, in which the goal is to provide an accurate, impartial, quantitative description of the findings in a population of studies on a given issue of interest. The survey may be done either by exhausting the population or sampling exhaustively from it. And of course, no survey would be considered valid if a sizable subset (or stratum) of the population was not represented in the cumulative results (Smith, 1980).

14.2 Mixture methods for discriminating between the linear quadratic and target theory models of cell-survival

We have noted above that since the target theory and linear quadratic models of cell survival, each with three parameters, are non-nested models, the methods of model discrimination based on the chi-squared approximation to the sampling distribution of decrements of deviance or chi-squared between the rival models are *not* applicable. Thus, we have used quality criteria such as AIC and PRESS, or posterior odds ratios, none of which are sampling measures.

We have also discussed the possibility of discriminating between these two models on the basis of additional experimental observations in the dose region around 100 rads, at which the respective predicted levels of response for the two models show the maximum difference. However, experiments in this region of dose are well-known to be difficult (Hall, 1975).

An alternative procedure for formally discriminating between two non-nested rival models that, unlike the AIC and PRESS statistics, or the posterior odds ratio, does depend on the sampling distribution of the discrimination criterion, requires that the two rival models, say M1 and M2, be *embedded* in a larger model, say M_0, which includes each as a special case, distinguished by characteristic values of an index parameter, λ. This has transformed the problem of *discrimination* into a problem of *estimation* (Gilchrist, 1984). For example, M_0 is a mixture model with index parameter λ:

$$f_0(\underline{x}_i^T,\ \underline{\beta}_0) = f_1(\underline{x}_i^T,\ \underline{\beta}_1)^{\lambda} f_2(\underline{x}_i^T,\ \underline{\beta}_2)^{1-\lambda}$$

where $\lambda = 1$ corresponds to the model M1 with form $f_1(\underline{x}_i^T,\ \underline{\beta}_1)$ and $\lambda = 0$ corresponds to the model M2 with form $f_2(\underline{x}_i^T,\ \underline{\beta}_2)$. Point and interval estimates of λ may be used to discriminate between M1 and M2. This procedure has been used to discriminate between the rival *absolute* and relative risk models of the incidence rate for radiation-induced leukemia (Muirhead and Darby, 1987). It could also be deployed to discriminate between the rival LQ and Target theory models of cell survival rates and this is the second response of the Task Group.

The mixture methods are not without characteristic weaknesses. One possible difficulty is that the minimum number of observations, say n, that are required is much larger than is usually found for most cell survival experiments; at least n = 20 are required and 30 would be preferable. Moreover, as Preston (1990) has remarked, for a mixture model of some data the interval estimates on λ may be too wide to provide an unambiguous identification of one or the other model. Moreover, interpretation of the model may be difficult unless λ is near either 0 or 1.

14.3 Bayesian hierarchical meta-analysis. "Mouse-to-man". Interspecies extrapolation of dose-response functions.

"The second comment that I have is that all non-experimental data sets are simply too weak to allow sensible inferences in the absence of supplementary information."

E. Leamer, 1986

"The extrapolation of the results from laboratory animal studies to man is probably the most difficult and least understood aspect of the total extrapolation problem."

P. Gehring, 1977

"... convincing estimates of human cancer risk from non-human cancer risk from non-human studies require evidence from both multiple [carcinogens] and multiple species, including at least some informative data in humans. Human risk extrapolations based solely on the results of one compound in one non-human species will be highly sensitive to prior beliefs about the relevance of the non-human species to man."

W. DuMouchel and J. Harris, 1983

"This effect, that one can improve upon the standard estimators of 'independent problems' through combined estimation, will be called the stein effect."

J. Berger, 1982

"The great benefit of the Bayesian approach to this problem is that it makes precise the assessments of relevance and uncertainty in related results."

R. Kass, 1983

The third response of the Task Group 1 to the question of what to do now requires the deployment of a recently developed class of Bayesian statistical methods - Bayesian hierarchical meta-analysis - for combining the information from several dose-response studies in order to reduce the variance of the estimates of the model parameters - and of the estimated response - for dose-response models of clinical radiation response data. Although this problem is also one in meta-analysis of a set of studies, it is complicated by the requirement that the dose-response studies to be integrated must be based on different species, since, in principle, dose-response studies in animal experiments should yield more precise estimates and inferences for a given dose-response model - or family of such models - than do clinical studies in patient populations since the investigator is less constrained in the choice of design parameters for animal experiments. Thus, to the now familiar problems of model construction, criticism, discrimination, validation, and deployment that were examined above is now added the most intractable - transportation. That is, the transport of dose-response models between two different species exposed to a common agent. Moreover, it must be noted, as the secondary analyses described in this report have disclosed repeatedly, that the animal studies also may - most unfortunately - frequently be encumbered by weaknesses as destructive of sound inference and estimation as are those found in cognate clinical studies. For example, it is too often the case that, too few animals are placed at risk to resolve the matter at issue in any given study. That is, there are far too few of what, in section 11, we have termed Type B4 studies.

DuMouchel and Harris (1983) have proposed, "... a class of Bayesian statistical methods for interspecies extrapolation of dose-response functions [Bayesian hierarchical meta-analysis]. The methods distinguish formally between the conventional sampling error within each dose-response experiment and a novel error of uncertain relevance between experiments. ... the dose-response data from many substances and species are used to estimate the inter-experimental error. The data, the estimated error of interspecies extrapolation, and prior biological information on the relations between species or between substances each contribute to the posterior densities of human dose-response."

The Bayesian hierarchical models provide a natural way of attacking the "mouse-to-man" problem. They impose a formal theoretical structure on the previously ill-defined problem of interspecies extrapolation. The methodology is quite general and has been used to construct both point and interval estimates of risk in humans following exposure to either radionuclides (DuMouchel and Groer, 1989) to chemical carcinogens (DuMouchel and Harris, 1983). Such models are erected on the assumption that the results obtained from studies on different species exposed to different toxic agents are somehow related and that the data include significant information on the relationship. The models use dose-response data from many different exposures and many different species to estimate the inter-experimental error, that is, the method does not require a priori assumptions that any of the animal data are relevant to humans; instead the degree to which the animal and human data are mutually consistent is estimates from the analysis.

These Bayesian methods must distinguish explicitly between the conventional sampling error within each dose-response experiment and, in the DuMouchel and Harris locution (1983), the "novel error of uncertain relevance" between experiments. It is the case, of course, that interspecies transport of dose-response functions is often encumbered by large, "errors of uncertain relevance". For example, Ames et al, 1987 report that of 392 chemicals [mutagens] "... tested in both rats and mice, 226 were carcinogens in at least one test, but 96 of these were positive in the mouse and negative in the rat or vice versa." This may be interpreted in the present context as evidence that in one species the slope parameter for dose exceeds zero, $\beta_1 > 0$, while in the other species $\beta_1 = 0$, or vice versa.

Often before the gathering of any current data in a given investigation, there is strong relevant prior information in some form available from past studies or from theory or expert judgments. The received method for formally combining prior, or non-sample, information with sample information on a given issue to produce estimates and/or predictions, together with the respective estimates of uncertainty, that fully incorporates both sources of information is Bayes' Theorem.

Combining information from a variety of sources (statistics, data sets, experiments, bodies of expert opinion) can produce sharper answers to 'questions that may be answered, less accurately and concisely, on the basis of each information source considered separately." ... Indeed, appropriate combinations of sample and non-sample information may even be used to produce answers to questions that otherwise have no defensible answer ..." DuMouchel and Harris (1983) developed a Bayesian hierarchical model through which they were able - by "borrowing strength" in a reasonable way across an ensemble of human and non-human studies on chemical carcinogenesis - to produce a plausible estimate of the carcinogenicity in humans of a type of diesel fuel for which there were no available human data.

In general, hierarchical models can be summarized as follows. (Casella, 9185; Raudenbush and Bryk, 1985; NAS/NRC, 1992). Assume that the ith member of an ensemble of k isomorphic, or parallel, studies produces a point estimate y_i of a study-specific parameter θ_i together with a sampling error of estimate ϵ_i with variance σ_i^2. That is, y_i is distributed Normally with expectation $E(y_i) = \theta_i$ and $Var(y_i) = Var(\epsilon_i) = \sigma_i^2$. Or $y_i = \theta_i + \epsilon_i$, where $E(y_i) = \theta_i$ and $\epsilon_i - N(0, \sigma_i^2)$. It is further assumed that $\theta_i = \mu + \delta_i$, where μ is a hyper-parameter common to all studies and $\psi - N(0, \tau^2)$. Under this model, the variation of each y_i has two components, or

$$y_i = \mu + \delta_i + \epsilon_i$$

and

$$Var(y_i) = (\tau^2 + \sigma_i^2)$$

The natural estimate of the overall mean is the weighted mean

$$\overline{\mu} = \Sigma w_i y_i$$

with weights

$$w_i = (\tau^2 + \sigma_i^2)^{-1}/\Sigma_j(\tau^2 + \sigma_j^2)^{-1}$$

Although the random effects model assumes that each y_i has its own study-specific parameter θ_i, the "best" estimate of θ_i (for a mean squared error loss function) is not y_i, but rather the "shrunken" estimate $\overline{\theta}_i$ where

$$\overline{\theta} = (y_i\tau^2 + \overline{\mu}\sigma_i^2)/(\tau^2 + \sigma_i^2)$$

That is, each estimate is "shrunk" toward the weighted estimate $\overline{\mu}$ of μ.

The Bayesian hierarchical model can be extended to include the effects of study level covariates described by the vector $\underline{\mathbf{x}}_i^T$. Then the model for y_i is now

$$y_i = \underline{\mathbf{x}}_i^T\underline{\boldsymbol{\beta}} + \varsigma_i + \varepsilon_i.$$

$\underline{\mathbf{x}}_i^T$ includes both treatment variables such as dose and prognostic factors such as tumor stage. $\underline{\boldsymbol{\beta}}$ is an unknown parameter vector. Moreover, a full Bayesian approach requires prior distributions for $\underline{\boldsymbol{\beta}}$ and τ^2. (See DuMouchel and Herbert, 1993).

There are two sources of error in such models: the conventional within-study error of measurement, ε_i, and a novel between-study error of uncertain relevance δ_i. The Bayesian hierarchical models will, of course, provide estimates of both. The estimate of the error of relevance provides a measure of the risk of interspecies extrapolation of the study results. In cases in which the assumptions are valid and the error of relevance is small, Bayesian hierarchical models will provide shorter interval estimates (i.e., confidence intervals) of parameters where there is observed data and finite interval estimates where there is no data. Also, the Bayesian model can help to decide which of many other unperformed experiments might provide the most information on a given issue.

It is also the case in most meta-analyses, such as those that may be undertaken under 14.1 above, that the goal is usually to achieve a general conclusion or consensus as to which of several competing hypotheses seem most supported by the preponderance of the evidence in the literature - both published and unpublished studies. However, the Bayesian meta-analysis is concerned more with estimation than with inference, or hypothesis testing. In particular, it is concerned with strengthening the parameter estimates of a dose-response model of clinical radiation effects. As DuMouchel (1990b) has observed, "The Bayesian model allows data from some studies to assist in the estimation of parameters characterizing the other studies. This notion of different studies 'borrowing strength' from each other is a crucial one. However, in most realistic problems the degree to which different studies are relevant to each other is somewhat uncertain. Hence all but the simplest hierarchical Bayesian models also use the data to help decide how related the different studies are. Depending on whether or not the estimates from each study accord with prior beliefs reflected in the model relating the different studies, the studies may or may not be able to borrow strength from each other." And, "Most importantly, the method does not require the assumption that the animal data is relevant to humans. Instead, the degree to which the animal and human data agree can be estimated from the analysis." Thus, the "Bayesian hierarchical models provide a natural way of attacking the 'mouse to man' problem. For example, DuMouchel and Harris (1983) developed a Bayesian hierarchical model through which they were able, by "borrowing strength" in a reasonable way across an ensemble of human and non-human studies on chemical carcinogenesis, to produce a plausible estimate of carcinogenicity in humans of a type of diesel fuel for which there were no available human data. One important achievement of these hierarchical Bayesian methods is that one is able to specify with, say, 0.95 credibility, the accuracy to be expected in the interspecies extrapolation - from animal experimental data to human clinical data of the dose-response information captured in say, the study-specific estimates of the parameter $\underline{\boldsymbol{\beta}}$.

The hierarchical Bayesian meta-analysis is characterized by the construction and deployment of a formal statistical model at two levels. First, a parametric model is constructed for each of the individual dose-response studies to be integrated. Each study is characterized by the parameter of the model. The parameters may be the slopes, say β, of dose-response curves, relative risks, say

RR, the ED_{10}'s, etc. Second, another parametric model, with parameter $\underline{\theta}$, is constructed which relates the characteristic parameter vectors of each study to one another. $\underline{\theta}$ may be referred to as a hyperparameter. The two levels of the hierarchical model are combined by Bayes formula for the conditional probability; the Bayesian model allows the component studies to "borrow strength" from one another.

To construct a Bayesian hierarchical model of the data from several different studies it is necessary that the studies have a certain commonality, or parallelism, that is, that the meanings of the respective parameters which summarize the data of the several studies are all related and commensurable, e.g., that all are slopes, $\hat{\beta}_j$, or are all relative risks, RR, etc. (Darby (1986) has made a similar observation anent parallel studies in radiation epidemiology: "As one step in the process of establishing whether it is appropriate to generalize from the experience of the studies described above to other populations exposed to radiation, it is useful to analyze data from the two [or more] studies using so far as is possible, identical methodology: and then to compare the findings. In this way any differences between the two studies will be highlighted, and if none are found the results of the studies will be in form suitable for combination."

The DuMouchel-Harris procedure also requires that each study-specific statistic, say $\hat{\beta}_j$ or RR, from the several individual, parallel, studies, include a measure of its uncertainty, e.g., $\hat{\beta}_j$ and $Var(\hat{\beta}_j)$, RR and Var(RR), etc. If the dose-response data within each study is multivariate and therefore must be represented by a matrix of treatment variables and covariates instead of a vector then the first level of the hierarchical model is represented by the set of sample estimates of parameter vectors $\hat{\underline{\beta}}$ and the respective covariance matrices $Var(\hat{\underline{\beta}})$. Since many published studies do not include interval estimates of the parameters, and moreover, may have mis-specified the model, mis-analyzed the data, etc., it may be necessary to re-analyze the data (secondary analysis) of many (all?) of the component studies. As Dumouchel has remarked, it "... is [a] fact that many published studies do not contain enough statistical detail to enable the parameter estimates and, especially, their standard errors, to be calculated. A particular author may have taken a different analysis tack than that required by the proposed meta-analysis, or even mis-analyzed the data. To construct the needed set of parallel parameter estimates with associated standard errors it may be necessary to re-analyze many of the original data sets. [emphasis added] This often requires the cooperation of the original researchers and involves much more time and effort than does the meta-analysis itself." (DuMouchel and Harris, 1983). Thus, as noted by both Darby and Dumouchel, it is usually the case that a secondary analysis of some or all of the studies to be integrated must precede any attempt at synthesis. It also appears to be the case that the failure to publish assessments of the uncertainties inherent in the results obtained in a given study, a weakness that we have described for the radiobiological literature, is a quite general failing in many other fields.

The primary data of the several, say q, studies on distinct species on, say m, risk factors to be combined are described by a q*m matrix, of study specific parameters, say the dose-response slopes β_{jk}, or relative risks, RR_{jk}, together with their respective dispersion measures, say $Var(\beta_{jk})$ or $Var(RR_{jk})$, $1 \leq j \leq q$; $1 \leq k \leq m$.

DuMouchel and Harris developed the Bayesian hierarchical meta-analysis in a study of chemical carcinogenesis. In this brief exposition we will keep to that context. Let $y_{ij} = f(\beta_{ij})$, or $f(RR_{ij})$, where f(.) is a monotonic function, (including the identity), say f(.) = log(.), and let c_{ij} be the cognate standard error. Let θ_{ij} be the true value of y_{ij}, i.e., $y_{ij} - N(\theta_{ij}, c_{ij}^2)$. It is assumed that $\theta_{ij} = \mu + \alpha_i + \gamma_j + \delta_{ij}$ where μ is an overall mean effect, α_i is a species-specific effect, γ_j is a carcinogen-specific effect and δ_{ij} is a species-carcinogen interaction effect, i.e., $\theta_{ij} - N(\mu + \alpha_i + \gamma_j, \sigma^2)$, where $\sigma^2 = Var(\delta_{ij})$ is a measure of how poorly the θ_{ij} fits this additive model. The μ, α_i, γ_j, δ_{ij}, and σ are hyper-parameters. The observed summary statistics are described as generated by the model $y_{ij} = \mu + \alpha_i + \gamma_j + \delta_{ij} + \epsilon_{ij}$. Then we have $y = X\underline{\beta} + \underline{\delta} + \underline{\epsilon}$ where $\underline{\theta} = X\underline{\beta} + \underline{\delta}$. $\underline{\beta}$ is a column vector of parameters and X is an appropriately chosen design matrix. The ϵ_{ij} describe the within-experiment errors; $\epsilon_{ij} - N(0, c_{ij}^2)$. The δ_{ij} describe the between-experiment errors, $\delta_{ij} - N(0, \sigma^2)$. σ is the standard deviation of the error of relevance. The hyper-parameters have prior distributions. Specifications of these prior distributions, together with the statistics y_{ij}, c_{ij}^2 comprise the initial information required for the Bayesian model. Thus, the uncertain relevance between experiments is formalized by the assumption of a common hypothetical process that generates the data in all k*m experiments. There is a trade-off between predictive bias (σ) and predictive efficiency (c_i) in the posterior estimates: One may wish to include less relevant studies in order to increase the precision of posterior estimates. (DuMouchel and Harris, 1983).

The available a priori information on carcinogen characteristics, species differences, disease processes, etc., is included in the analysis in the form of prior information on the parameter of the underlying model that relates the set of different experiments. Posterior estimates of $\underline{\theta}$ and ($\hat{\underline{\theta}}^*$ and $\hat{\sigma}^*$) are obtained from the data and a priori information by means of Bayes' theorem.

...

DuMouchel and Harris, (1983) further remarked that, "Because of the statistical nature of the technique [Bayesian hierarchical meta-analysis], in the absence of strong prior information, many studies are needed in order to make useful interspecies extrapolations. Sometimes the lack of a critical study is important .. We need to think more in terms of ensembles of studies and how each new study fits into the pattern of previous studies." [emphasis added]

In their first paper, in which their highly original views on the general problem of combining diverse forms of scientific evidence on a given issue were presented, DuMouchel and Harris (1983) use chemical carcinogenesis data to point out the common statistical and biological problems encountered in any inter-species extrapolation. Their closing remarks, although addressing human cancer risk estimates, are generally applicable to any dose-response process: "... convincing estimates of human cancer risk from non-human studies requires evidence both from multiple substances and multiple species, including at least some informative data in humans. Human risk extrapolations based solely on the results of one compound in one non-human species will be highly sensitive to prior beliefs about the relevance of the non-human species to man ... when a compound's human cancer risk is predicted solely from experiments with that compound in a single non-human species such as rats (in effect a 2x1 table with an empty cell), then the resulting human risk estimates will depend entirely on one's prior beliefs about the conversion factor between rats and humans or about the relevance of rats to man ... Human cancer risk assessment requires data on many agents in many species. In the absence of strong prior information on cancer mechanisms, one good rat study is just not enough." [emphasis added].

In order to exploit the DuMouchel-Harris Bayesian hierarchical model in interspecies extrapolation of radiation dose-response information it is necessary to identify the chemical agent or compound of their cancer risk study with the LET of ionizing radiation. Let us therefore examine briefly the possible deployment of a priori information on $\underline{\beta}$ obtained from isomorphic animal studies, in the construction of dose-response functions, $\eta_i = \underline{x}_i^T\underline{\beta}$, for interpretation and extrapolation of non-experimental human data. This has already been done in the case of low dose effects - carcinogenesis - in humans (BEIR IV, 1989; DuMouchel and Groer, 1989). We recall that low dose effects in both experimental and non-experimental studies are typically described by dose-response models rather than the isoeffect models that typify high dose effects.

All of the information on the dose-response functions for humans exposed to therapeutic levels of low LET ionizing radiation is contained in sets of non-experimental observations. For some sets of these data, one (or more) of the polynomial (Taylor series) models may provide an adequate "fit" to the data, on the evidence of a single aggregate statistic such as the Pearson chi-squared. But this evidence implies only that a "significantly large" amount of the observed variation in the response, about the mean, is accounted for by the model. Ancillary regression diagnostics are required to further determine whether the model is "adequate," i.e., whether the model maps the observations into a white noise sequence and whether or not the estimate, $\hat{\underline{\beta}}$, is dominated by a subset of the data. A "significant" degree of fit does not invariably mean, however, that the fitted equation gives either the "true" description of process generating the observations or even that it is a worthwhile predictor of response in the Box-Wetz sense (Box and Wetz, 1973) that the range of predicted values, $\underline{x}_i^T\hat{\underline{\beta}}$, $1 \leq i \leq n$, is "substantial" compared with the standard error of the estimated response. (The test of significance criterion can only decide whether the model is better than the mean response as a predictor of response.) Even for statistically adequate models of "good" data the ratio of the parameter estimates to their standard errors, $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$ of the dose-response models are usually small, say 2-3, whereas for a model to provide a good predictive performance these ratios should be much greater, say 8-9.

Moreover, the risk manager, "who requires wisdom, as well as a forecast", from any estimator of $\underline{\beta}$ will usually seek to interpret the estimate, $\hat{\underline{\beta}}$. And, it is a common misapplication of regression methods to attempt to interpret an inadequately estimated response function. Thus, in order for the estimated model to be regarded as useful (as well as "adequate") the range of response values predicted by the equation, $|\underline{x}_{(1)}{}^T\underline{\beta} - \underline{x}_{(n)}{}^T\underline{\beta}|$, should exceed, say, ten times the (average) standard error of the estimated response, $|Var(\underline{x}_i{}^T\hat{\underline{\beta}})$, $1 \leq i \leq n$.

It seems to be the case that none of the probit (or logit) linear regression models estimated from the clinical studies is "worthwhile" in this more restrictive, Box-Wetz, sense. Therefore, it would seem to be desirable (necessary?) to "strengthen" the sample information on $\underline{\beta}$, obtained from such studies, with non-sample, or a priori, information on $\underline{\beta}$ - or on the response, $\underline{x}^T\underline{\beta}$ - obtained in isomorphic studies. A standard (quasi-) Bayesian maneuver for implementing this requirement is the Theil-Goldberger Mixed estimation methodology. (Theil, 1971; Theil and Goldberger, 1961). There are, of course, other maneuvers but this seems to be appropriate - and quite accessible. By isomorphic studies we mean that the respective sets of observations in the studies cover the same ranges of the same explanatory variables (or that an appropriate scaling is valid) and that the models of dose-response have the same form; that is, that they are generalized linear models of the same family, i.e., probit, Poisson, etc. Thus, the models of cell-survival data (Poisson response) and the models of clinical dose-response data (Binomial response) are not isomorphic. (We have remarked earlier that this is one of the conceptual weaknesses of the multifraction LQ model of radiation toxicity.)

We noted above that although the professional concern of radiation therapy is in maximizing the probability of uncomplicated control of disease there is at present neither experimental nor non-experimental data from which models of the joint probability of the concomitant occurrence of two or more radiation effects - in tumor and normal tissue - can be constructed. Nor, for that matter, does it appear that much thought been given yet to the computational methods for estimation of the parameters - which must include the correlation coefficients that describe the association of the tolerances of the two systems, e.g., normal and tumor tissues - of these models. (See Herbert, 1993b; Lesaffre and Molenberghs, 1991). Obviously, there is a considerable amount of work that must be done in both the clinical and laboratory studies before the information obtained from the models of the latter can be deployed to stabilize the estimates of the parameters in the (isomorphic) models of the former.

It will be most useful, at this point, by way of developing a realizing sense of some of the essential issues, to examine in some detail a rather more relevant, as well as more familiar, example of an inter-species extrapolation of radiation dose-response information. This is found in the cancer risk estimates of the BEIR III Report (NAS/NRC, 1980) in which the weaker sample evidence on the received models of radiation dose-response of the LSS non-leukemia cancer mortality data is strengthened by combining it with cognate stronger evidence of the LSS leukemia incidence. See Fig. 38a. But it should be noted that our example is an instance of Mixed estimation, not of a hierarchical meta-analysis such as described by DuMouchel and Harris (1983) and DuMouchel and Groer (1989). It is, however, an example of an interspecies extrapolation, or transfer of dose-response functions, since leukemia and cancer are different species of tumor: one is an epithelial tumor, the other a non-epithelial tumor. The respective natural histories, incidence rates, and radiation responses are quite different. Peto (1977, 1979) and Muirhead and Darby (1987) treat the epidemiological data on these two classes (species?) of tumours as being mutually incompatible (or uninformative).

The leukemia information on the parameter vectors of the models of non-leukemia cancer mortality data, can be represented by the linear constraint, $\underline{r} = R\underline{\beta} + \underline{v}$, $Var(\underline{v}) = \psi = [0]$, on the sample estimates, $\hat{\underline{\beta}}$, of the respective LQ-L, L-L, Q-L models (See Annex IV, part 5 and Herbert, 1986c). The elements of the matrices $\underline{r}$ and R, are functions, $\theta = f(\underline{\beta})$, of the estimated parameter vectors, $\hat{\underline{\beta}}$, for the cognate Poisson linear models of the leukemia incidence data. Since the range of observations was the same in both the leukemia incidence and non-leukemia cancer mortality data, and the respective dose-response models, LQ-L, L-L and Q-L were Poisson linear regression models, these two studies were isomorphic.

Figure 38a presents an index plot of the leukemia incidence and cancer (sans leukemia) mortality rates for the BEIR III (1980) LSS data where H and N denote the Hiroshima and Nagasaki samples, respectively.

Figure 37 describes the pooling of the Hiroshima and Nagasaki samples of leukemia incidence data. This salvage maneuver (Data augmentation) was successful in achieving a lower degree of correlation between D_γ and D_n in the pooled sample than in the respective city-specific samples because the correlation structures of the latter are complementary (Herbert, 1986c). More acceptable sample estimates of the parameter vector $\underline{\beta}$ of the LQ-L, L-L, and Q-L models of leukemia incidence were obtained from the pooled data.

This maneuver also reduced the degree of correlation of D_γ and D_n in the cancer (sans leukemia) mortality data, of course. However, more acceptable sample estimates of the parameter vector, $\underline{\beta}$, of the dose-response models LQ-L, L-L, and Q-L still could not be obtained from those data. This is due to the fact that, since the response has a (conditional) Poisson distribution for which the variance is equal to the expectation, the (higher) mortality rates are noisier than are the (lower) leukemia incidence rates.

Therefore, the linear constraints, $\underline{r} = R\underline{\beta} + \underline{v}$, $E(\underline{v}) = \underline{0}$, $Var(\underline{v}) = \psi$, were imposed on the sample estimates of $\underline{\beta}$ for these models of the pooled mortality data in order to increase the precision of the sample estimate of $\underline{\beta}$. The elements of the matrices, $\underline{r}$ and R were obtained from the point estimates of $\underline{\beta}$ for the cognate models of the LSS leukemia incidence data. However, it was stipulated by the BEIR III Committee that $\psi = [0]$, the null matrix (although it is evident from Table V-8 of the BEIR III (1980) report that this stipulation greatly over-states the precision of the (pooled) LSS sample estimates of $\underline{\beta}$ for those models of the leukemia incidence data. (For example, in Table V-8 the sample estimates of β_1 and β_2 barely exceed the respective standard errors.)

The BEIR III estimates of $\underline{\beta}$ for the LQ-L, L-L, and Q-L models of cancer sans leukemia mortality were obtained as posterior estimates by using the procedure of Mixed estimation. For a Normal theory model we have

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \underline{\beta} + \begin{pmatrix} \varepsilon \\ \underline{v} \end{pmatrix}$$

This can be generalized to the Poisson models by the transformations,

$$P\underline{y} = PX\underline{\beta} + P\underline{\varepsilon}$$

and

$$Q\underline{r} = QR\underline{\beta} + Q\underline{\varepsilon}$$

where $PP^T = V^{-1}$ and $QQ^T = \psi^{-1}$ where $V^{-1} = W$ is the (n*n) Poisson weight matrix of the sample data. The proportion of a priori, or non-sample, information in the posterior estimates of $\underline{\beta}$ in Table V-11 for cancer (sans leukemia) mortality that is contributed by the leukemia incidence data of Table V-8 can be shown to be $p = 0.41$ (See Herbert, 1986b, 1989d).

It is obvious that a similar salvage maneuver - Mixed estimation - can be deployed to obtain more precise estimates of the parameter vector $\underline{\beta}$ for models of radiation toxicity in which the response has a Binomial distribution just as in the present case in which the response has a Poisson distribution.

Figure 38b presents an index plot of the standardized residuals for the LQ-L model of the BEIR III LSS (1980) cancer (sans leukemia) mortality data obtained by Mixed estimation:

$$\begin{pmatrix} P\underline{y} \\ Q\underline{r} \end{pmatrix} = \begin{pmatrix} PX \\ QR \end{pmatrix} \underline{\beta} + \begin{pmatrix} P\underline{\varepsilon} \\ Q\underline{v} \end{pmatrix}$$

The two pseudo-observations at i = 17, 18 represent the leukemia incidence information on the

LQ-L model that is conveyed by the constraint $\underline{r} = R\underline{\beta} + \underline{v}$; $E(v) = 0$, $Var(\underline{v}) = \psi = [0]$, where $\underline{r}$ is (2x1), R is (2xk), etc. The matrices P and Q are weight matrices appropriate to the cancer mortality (sans leukemia) sample data and the leukemia constraint, respectively: $P^TP = V^{-1}$ where V^{-1} is the (nxn) Poisson weight matrix of the sample obtained from the IRLS estimates of $\underline{\beta}$ for the LQ-L model of non-leukemia cancer mortality and $Q^TQ = \psi^{-1}$ is the (2x2) weight matrix of the leukemia constraint. See Belsley et al, 1980 and Herbert, 1986b, 1989d.

Note that the two pseudo-observations, which represent the leukemia incidence sample, dominate the estimates of $\underline{\beta}$ for the LQ-L model of radiation carcinogenesis (sans leukemia). This plot should be compared with those for the LQ models of the Tradescantia radiation mutagenesis data in Figs. 5b, 6b, and 8 in which the observations in the sample of experiment #6 dominate the estimate of $\underline{\beta}$ in the pooled sample which includes the data of experiments #2, #5, #6, and #7.

It should be remarked that the use of a priori information on $\underline{\beta}$ that is obtained from the leukemia incidence data to stabilize the parameter estimates of the LQ-L, etc., model of cancer (sans leukemia) mortality data represents an instance of the "interspecies transfer of dose-response functions". But, "While the mechanisms of induction of most carcinomas may all be rather similar, they probably differ fundamentally from the mechanisms of induction of leukemias, sarcomas, etc., and much of the research into leukemia viruses, or sarcoma viruses may be irrelevant to the 90% of human tumors that are carcinomas" (Peto, 1977).

Figure 38c presents an index plot of the hat matrix diagonals for the LQ-L model of the BEIR III (1980) LSS cancer (sans leukemia) mortality data obtained by Mixed estimation. It is obvious that the two pseudo-observations representing the leukemia incidence information completely determine two of the five components, β_j, of the parameter vector, $\underline{\beta}$. This information from the diagnostics is consistent with the estimate $\hat{\theta}_p = 0.40$ obtained from the Mixed estimation technique. See Belsley et al, 1980 and Herbert, 1986b, 1989d.

Moreover, the other single-row deletion diagnostics DFBETAS, COVRATIO and DFFITS, as well as PRESS, are functions of the two key diagnostics, e_i^* and h_i. In general, these diagnostics will be large when either e_i^* or h_i - or both - are large. Thus, it appears that the sample estimates, $\hat{y}$, $\hat{\underline{\beta}}$, $Var(\hat{\beta})$ and PRESS for the LQ-L model of non-leukemia cancer mortality are dominated by the pseudo-observations #17 and #18 which represent the information on $\underline{\beta}$ obtained from the sample estimates of the leukemia incidence rate. This would seem to raise some issues of interpretability of the posterior estimate $\hat{\underline{\beta}}^{**}$ of the parameter vector $\underline{\beta}$ in the LQ-L model on non-leukemia cancer mortality.

Of course, the relevance, as well as the compatibility (which is measured by the statistic, γ, vide supra, of the non-sample information, $\underline{r} = R\underline{\beta} + \underline{v}$, must be examined and included in any account. Thus, although on the evidence of the Mixed estimation concordance γ statistic (see section 7.2.4), the (a priori) leukemia incidence information on $\underline{\beta}$ is not incompatible with the non-leukemia mortality sample information, its relevance is surely questionable. (Non-leukemia cancers are epithelial tumors. Leukemias are non-epithelial tumors.) One of the unique strengths of the DuMouchel-Harris (1983) procedure is, of course, that it provides an empirical measure of the "... novel error of uncertain relevance" and of the trade-off between predictive bias and predictive efficiency." As noted above, "... the method does not require the assumption that the animal data is relevant to humans. Instead, the degree to which the animal and human data agree can be estimated from the analysis." (DuMouchel and Harris, 1983).

In place of a priori evidence from other non-experimental human studies, as was used in BEIR III, one may consider obtaining the a priori information on $\underline{\beta}$, represented by the constraint $\underline{r} = R\underline{\beta} + \underline{v}$, from animal (rat, mouse, etc.) studies that are isomorphic, as in the Bayesian methods of mouse-to-man extrapolation developed by DuMouchel and Harris (1987) that were described above. Note also that, in this case, since the posterior estimate, $\hat{\underline{\beta}}^{**}$, of $\hat{\underline{\beta}}$ is a matrix-weighted average of the information on radiation dose-response of two distinct animal species, it might properly be referred to as a chimerical estimate. But since the second meaning of that term seems a bit pejorative, we choose to avoid it.

The BEIR V (1990) report subdivided the tumors in the LSS sample according to organ system instead of using the BEIR III (1980) dichotomy of leukemia and non-leukemia cancer, since the additional years of follow-up had strengthened the data. As remarked above, the BEIR V analysis showed that the mortality rates for all tumors other than leukemia had a <u>linear</u> rather than a linear-quadratic dependence on radiation dose. Figure 38d is a plot of the Freeman-Tukey residuals,g_i, vs the estimates response $\hat{\mu}_i$ for the linear model of the digestive tumors in the LSS sample (DS86) dosimetry. The Freeman-Tukey residuals are the preferred measures of concordance because they are more <u>robust</u> than either the deviance residuals, d_i, or the Pearson chi-square residuals, χ_i when the expected Poisson response, $\hat{\mu}_i$, is small: d_i and χ_i are strongly <u>inflated</u> as $\hat{\mu}_i$ 0, since each is a function of μ_i^{-1} and are <u>undefined</u> at $\hat{\mu}_i = 0$. Moreover, as $\hat{\mu}_i$ 0, the respective sampling distributions for $\Sigma\chi_i^2$ and Σd_i^2 are not well-approximated by the Pearson chi-squared distribution with degrees of freedom (n-k) for a k parameter model of a sample of size n. Since the BEIR V data are <u>disaggregated</u>, with many cells with $\hat{\mu}_i \simeq 0$, the Freeman-Tukey residuals are preferred (<u>N.B.</u>: It will be useful to recall the definitions of these three residuals for Poisson data: $g_i = \sqrt{y_i} + \sqrt{y_{i+1}} - 4\sqrt{\mu_i+1}$; $d_i = [2y_i \ln(y_i/\hat{\mu}_i) - (y_i-\hat{\mu}_i)]^{1/2}$; $\chi_i = (y_i-\hat{\mu}_i)/(\sqrt{\hat{\mu}_i})$. Note that in Fig. 38d, g_i is plotted against $2\sqrt{\hat{\mu}_i}$, the constant information scale for a Poisson error distribution. It can be seen that the first few contours of constant y_i have (approximately) a slope of -1 (McCullagh and Nelder, 1983 and 1989). Compare with Fig. 18c, the cognate plot for a Binomial error distribution. Fig. 38e is a plot of y_i vs $\hat{\mu}_i$ for the linear (in dose) model of the BEIR V mortality data for digestive tumors. It can be seen that $y_i \simeq \hat{\mu}_i$ to a good approximation. Note that the plot includes 2162 records in 726 of which we have $y_i \geq 1$. (NAS/NRC, 1990).

<u>Important Topics</u>

Ethics of scientific enquiry. Esoteric science. Exoteric science. Fleck. Wolin effect. Law of Small Numbers. Zipf's Principle of Least Effort. Curvature of a plane curve. Law of the Instrument. Modus tollens. Lay and expert witnesses. Adversarial vs inquisitorial systems. Toxic tort. Classical meta-analysis. Bayesian hierarchical meta-analysis. "Mouse-to-man problem." Interspecies transfer of dose-response functions. Hyper-parameters.

14.4 <u>Multivariate probit analysis.</u>

There is yet another fruitful response to the fourth of the initial question proposed by TG1 (see section 1.4 above): Multivariate probit analysis. The multivariate probit model is a regression of a <u>vector</u> of correlated quantal responses on a set of continuous and/or discrete predictor variables. It fully exploits the correlation structure of the response and thus provides a solution to the multivariate response problem, which was introduced in section 6.1.1, and for which a heuristic discussion is presented in Appendix II. Several applications of the multivariate probit model can be found in the biological, economic, and psychosociological literature but it has not hitherto been widely used in biomedical studies.

Software for the multivariate probit model is not yet widely available. However, it is expected that this impediment will soon be removed (See Lesaffre and Molenberghs, 1991 and Steinberg, 1989). Therefore, the application of this model in both clinical and animal studies of the joint occurrence of characteristic radiation responses in normal and tumor tissues should prove to be as fruitful as it is now feasible. See Herbert, 1993b.

15. <u>Acknowledgments.</u>

"If asked to name varieties of mental torture, most scientists would place writing at the top of the list."

A. Kornberg, 1950 Nobel Laureate

"Scientific knowledge, like language, is the common property of a group or else it is nothing at all."

T. Kuhn, 1962

We wish to acknowledge, with thanks, his sustained and keen interest in, and his generous support of, the work described in this report by Professor B.G. Brogdon, M.D., University Distinguished Professor and Chairman of the Department of Radiology, College of Medicine, University of South Alabama. We also wish to acknowledge, again with special thanks, the patience, dedication, and expertise of Ms. M. Woodall, B.A., who cheerfully put up with and skillfully accomplished the steady rewrites and changes in the preparation of the manuscript. We also wish to acknowledge, again with thanks, the assistance - as well as the forbearance - of the members of TG1 in the review and the revision of this report. We also wish to acknowledge the contributions of several distinguished scientists who read and commented on various drafts of the report: A. Brodsky, Sc.D., J. Fowler, Ph.D., M.D. (Hon.), D. Jette, Ph.D., H. M. Kuan, Ph.D., L. Lanzl, Ph.D., M. Schell, Ph.D., P. Steward, Ph.D., L. Wagner, Ph.D., D. Zellmer, Ph.D. We also wish to acknowledge the assistance of the AAPM SC: M. Astrahan, Ph.D., M. Bronskill, Ph.D., E. Chaney, Ph.D., C. Coffey, II, Ph.D., C. Kelsey, Ph.D., E. Krishnan, M.D., P. Lin, Ph.D., C. Ling, Ph.D., R. Nath, Ph.D., R. Price, Ph.D., D. Ragan, Ph.D., T. Schultheiss, Ph.D., G. Simmons, Ph.D., S. Thomas, Ph.D., D. Tolbert, Ph.D., R. Wilson, Ph.D., and J. Zagzebski, Ph.D. In addition, we thank A. Shah, Ph.D., Professor of Statistics, University of South Alabama for carefully reading the entire report in its final form. Of course, the responsibility for any errors or any infelicities of thought and expression that may remain lies entirely with the principal author.

16. Epilogue.

"The justification sometimes advanced that a multiple regression analysis on observational data can be relied upon if there is an adequate theoretical background is utterly specious and disregards the unlimited capacity of the human intellect for producing plausible explanations by the car-load lot."

K. Brownlee, 1967

"The proposition that some inherent logical incompetence attaches to an inference based on observational, as distinguished from experimental, evidence seems to have little to commend it beyond the great positiveness with which it is asserted."

J. Cornfield, 1984

"We must, however, recognize that the fitting of equations to observational data (as opposed to data from carefully designed experiments) is, at best, a risky business."

R. Hocking, 1983

"... there is no substitute for a careful, or even a meticulous, examination of all original papers purporting to establish new facts."

R. A. Fisher, 1936

"When examining normal science ... we shall want finally to describe that research as a strenuous and devoted attempt to force nature into the conceptual boxes supplied by professional education."

T. Kuhn, 1962

"It isn't that they can't see the solution. It is that they can't see the problem."

G. K. Chesterton, 1923

"But when a compound's human cancer risk is predicted solely from experiments with that compound in a single non-human species such as rats (in effect, a 2 x 1 table with an empty cell), then the resulting human risk estimates will depend entirely on one's prior beliefs about the conversion factor between rats and humans or about the relevance of rat to man."

W. DuMouchel and J. Harris, 1983

"All nonexperimental data sets are simply too weak to allow sensible inferences in the absence of supplementary information."

E. Leamer, 1984

"If the results go unchallenged the researcher(s) involved may use the same substandard statistical methods again in subsequent work and others may copy them."

D. Altman, 1982

"As one step in the process of establishing whether it is appropriate to generalize from the experience of the studies described above to other populations exposed to radiation, it is useful to analyze data from the two [or more] studies using so far as is possible, identical methodology; and then to compare the findings. In this way any differences between the two studies will be highlighted, and if none are found the results of the studies will be in suitable form for combination."

S. Darby, 1986

16.1 Introduction.

" For every complex question, there is a simple answer - and it is wrong."

H. L. Mencken, 1927

Following the completion of the sections 1-14 of the present manuscript and before the review of the final draft of these sections by the Task Group 1 was completed, we had, as remarked in section 14.1, begun the review and evaluation of two recent and related studies in the inter-species extrapolation of dose-response functions, i.e., the "mouse-to-man" problem. These were the studies of Mah et al (1987) and Travis and Tucker (1987). As noted above, since the Travis and Tucker report currently appears to provide much of the received empirical evidence for the (received) form of the time factor, γT, for the multifraction LQ model, i.e., the "LQ + time" model, the validation of the findings of this report is therefore of no small interest. (For example, "The present form of the equation appeared most clearly in an editorial by Travis and Tucker (1987) ..." Fowler, 1989. See also Yaes, 1987). Therefore, we have decided to include the results of our review and secondary analyses of these studies in an epilogue to the report. Travis and Tucker 1987 has been cited 39 times to March 1992.

16.2 Summary of the Mah et al and Travis and Tucker studies

Mah et al used the information on the exponents of N and T obtained from the Wara et al mouse isoeffect experiments to construct an equivalent single dose, ED, for a simple probit dose-response model of a binary response in a sample of 54 radiotherapy patients. In both the animal and clinical observations the end-point was radiation pneumonitis. Travis and Tucker used the data from the mouse isoeffect experiments of Wara et al and Field et al (1977) to obtain an estimate of the coefficient of T in an LQ + time model to construct a rival equivalent single dose, ESD, for a simple logit dose-response model of the patient data in the Mah et al study.

These two 1987 studies were based upon, and closely followed in motivation, design, and analysis, an earlier study, that of Wara et al (1973) in which the information on the exponents of N and T obtained from isoeffect experiments in the mouse were used to construct a simple probit dose-response model of a binary response in a sample of 42 radiotherapy patients as did Mah et al in a sample of 54 patients. In both the animal and clinical observations the end-point was radiation pneumonitis. We note that although the Mah et al and Travis and Tucker studies repeated all of the more important mistakes of the Wara et al study of fourteen years earlier, they also included several additional (albeit not all of them original) mistakes in their estimates and inferences based on the Mah et al data.

16.3 "What do we believe?"

The several beliefs concerning the central matters of estimate and inference that are at issue in the pod of papers by Wara et al, Mah et al, and Travis and Tucker are fairly well rendered by the remarks quoted below:

"Naturally, the confidence in this slope [used in estimation of ED] is limited since its line was derived from only two points." [Their disclaimer notwithstanding, Ware et al used "this slope" in their subsequent calculations in 1973. So did Mah et al 14 years later.]

...

"The value of reducing treatment regimens to equivalent single doses is generally recognized."

...

"We urge caution in these dose ranges [2000 to 2500 rad in 3 weeks for whole lung radiation] and suggest a regimen yielding no more than 5% probability of pneumonitis for whole lung irradiation."

Wara et al, 1973

"The purpose of this work was to establish accurate dose-response data on acute radiation-induced pulmonary damage caused by fractionated radiation therapy."

...

"The goodness-of-fit to the data points was greater than 95% according to the chi-square."

...

"The findings of this study have generated a well-defined dose-response relationship between the incidence of acute radiation-induced pulmonary damage as observed on CT and the estimated single dose representation of fractionated radiotherapy schedules using the average lung dose in the high dose region."

...

"Figure 5 indicates that an average lung dose of 24.7 Gy in 15 daily fractions would give only a 5% incidence of pulmonary damage whereas 43.5 Gy in 15 daily fractions would produce acute pulmonary damage in over 95% of the patients receiving such a dose."

...

"In summary, this prospective study [by Mah et al, 1987] has established a distinct dose-response curve for human pulmonary tissues to fractionated radiotherapy."

...

"It is clinical data, collected in this prospective manner, that are essential for the clinical optimization of fractionation schedules, ..., and the verification of radiobiological models."

...

"Acute radiation-induced pulmonary damage, often referred to as radiation pneumonitis, can be a significant cause of morbidity and mortality."

Mah et al, 1987

"In all aspects, this study [by Mah et al, 1987] represents a carefully designed, controlled, executed and documented clinical trial."

...

"We are not proposing yet another isoeffect formula. The purpose of the analysis described above (i.e., the LQ + time model) was simply to determine whether the fit of the Wara model to the data of Mah et al could be explained in terms of cell survival as appears to be the case as shown in Fig. 2."

...

"We are not proposing a new isoeffect model. Rather the LQ + time model was used to show that for the treatment schedules represented in the study of Mah et al the empirical Wara model is consistent with the concept that tissue response in lung is determined by the extent of cell-killing."

...

"Thus, the Wara formula might be useful for representing different fractionation schedules with the proviso that it be applied only to conventional regimens using only one fraction daily."

Travis and Tucker, 1987

"The present form of the equation [of the LQ + time model] appeared most clearly in an editorial by Travis and Tucker (1987) ..."

J. Fowler, 1989

All three studies, Mah et al, Travis and Tucker, and Wara et al, appear to be quite influential; they have been widely cited since their publication: 13 times, 17 times, and 166 times, respectively (to June 1990). Each describes the construction of a univariate dose-response model of a sample of binomial clinical data on radiation pneumonitis in which ancillary non-sample information derived from animal (mouse) isoeffect experiments in radiation pneumonitis is used to transform the set of clinical treatment regimens (dose,D, fractions,N, and elapsed time,T) of the sample to "equivalent single doses", i.e., the level of dose, say D^*, given in a single treatment, N=T=1, that would elicit the same level of the same stochastic response as the actual clinical treatment regimen (D,N,T) where N >> 1, T >> 1. It must be remarked at once that in none of the three studies was the more appropriate probit (or logit) <u>matrix model</u>, say $z_i = \underline{x}_i^T \underline{\beta}$, $1 \leq i \leq n$, where the treatment vector $\underline{x}_i^T$ is (1*k), $k \geq 3$, of the <u>multivariate</u> clinical dose response data that was available, constructed and evaluated.

The several motives that were given by the respective investigators for the deployment of the isoeffect information obtained from the animal experiments in the dose-response models of clinical data for the three reports are both vague and various:

1) Wara et al. "The problem,then, is to determine which of these formulae [$D = NSD*N^{0.24}T^{0.11}$ derived from isoeffect clinical data on skin response vs $D = ED*N^{0.377}*T^{0.058}$ derived from isoeffect mouse data on acute pneumonitis] best fits the clinical data available for radiation pneumonitis in the human patient." ... "The value of reducing treatment regimens to equivalent single doses is generally recognized." [But no statistically adequate measures and criteria of goodness-of-fit or model discrimination and neither reasons nor references to support their conclusion on the putative value of equivalent single doses are given by the authors.]

2) Mah et al. "Available fractionation schedules were limited. Therefore, a model was necessary to represent those fractionation schedules with equivalent biological effects. The estimated single dose (ED) model of Wara et al was selected. It accounts for the relative contributions of total dose, fraction numbers, and treatment time to the effective dose."

"Because of clinical limitations of the current study, a model of effective single dose was necessary for the comparison of fractionation schedules" ... "The ED was used because it was derived from the response of normal lung tissues to fractionated radiation albeit in mice and was thus the best model available." ... "In contrast to the nominal standard dose, NSD, the ED can also be empirically extrapolated to a single dose as demonstrated by the similarities between the ED values for the LD50 from 2 to 20 fractions and the single fraction LD50 in mice." [See Wara et al Table <u>1</u> for mice. The issue of the possible irrelevance of animal isoeffect experiments to clinical practice is not addressed by Mah et al. That is, the estimates of the exponents of N and T obtained 14 years earlier from the Wara et al mouse experiments are applied with apodictic assurance by Mah et al to the clinical dose-response data in their construction of the ED.]

"Although animal models have provided much radiobiological insight, extrapolation of quantitative response data from animals to humans is uncertain." [However, the authors make no effort either to describe or to assess the nature and degree of this uncertainty and its possible effects on the Wara model of their data. And although the available fractionation schedules <u>are</u> limited, the authors do not indicate why this weakness of the clinical data can justify the use of an effective single dose model.]

3) Travis and Tucker. "However, there is an urgent need to accurately and concisely define isoeffect relationships for human normal tissues <u>in vivo</u> because of the implied steepness of dose-response curves as derived from animal data." [But no references to support this conclusion are given. Moreover, the remark is something of a non-sequitur: The unfortunate fact that, "isoeffect relationships" for a single level of response can convey very little information on the "steepness of dose-response curves" seems to have been over-looked by many. We shall examine this in more detail below.]

"The fact that the purely empirical Wara model provided a better description of the data than the CPK model is surprising since Cohen's CPK model relates tissue tolerance to cell survival. This discrepancy led us then to investigate whether the data of Mah et al could in fact be explained in terms of cell survival in the lung."

"The purpose of the analysis described above (i.e., the LQ + time model) was simply to determine whether the fit of the Wara model to the data of Mah et al could be explained in terms of cell survival as appears to be the case as shown in Fig. 2" [Figure 2 of the Travis and Tucker 1987 report is the logit dose-response curve on the ESD. It is, of course, not unlikely that morbid conditions in several cell populations and tissue are required for radiation pneumonitis to result; it seems a bit overweening to base the model wholly on the occurrence of lethal events in a single unidentified cell population.]

16.4 Narrative review of the studies of Wara et al, Mah et al, and Travis and Tucker.

We describe first the findings from our preliminary evaluation of these three studies and then we present the results of the respective secondary analyses. We here note that these findings and results do provide a kind of cross-validation of nearly all of the more important conclusions on the received data analytic practices in radiation biology that were presented in Sections 1-14 of the present report, as well as in Appendices I and II and Annexes I-IV, since one or more instances of each of the several statistical solecisms and scientific follies described and illustrated therein, together with the consequent errors of estimate and inference entailed thereby, could be identified in the papers of Mah et al and Travis and Tucker - as well as in those of Wara et al. The persistence in the more recent literature of the ontological and epistemological weaknesses which we had found in the earlier literature was confirmed. In several instances this represents a "propagation of error" - a variety of the intellectual Bourbonism which we have remarked before in other studies - over several decades of scientific practice.

1) We note first that although each of the three papers is concerned with the construction, evaluation, and clinical deployment of dose-response models all three fail to provide: a) Statistically adequate measures (aggregate and case statistics) of the goodness-of-fit of the respective models that were deployed to describe their data. [Although the Wara et al paper states that one of its objectives is to determine "... which of those formulae best fits the clinical data available ..." the measures and criteria of concordance of their rival (NSD vs ED) clinical models of dose-response which they present are wholly inadequate.] In their evaluation of the goodness-of-fit of their isoeffect models, Wara et al report that they have a, "regression coefficient of 0.999" and, "a regression coefficient of 1". [It appears that they misspoke themselves; "correlation coefficient" would be correct; a regression coefficient is a "slope"] The Mah et al report does proffer the rather bizarre locution that, for their probit model of their clinical data, "The goodness of fit to the data points was greater than 95% according to the chi-squared"! (Almost any reasonable interpretation of this rather garbled statement suggests that the Mah et al model does not fit their data. The percentages that are associated with goodness-of-fit statistics, such as "the chi-squared", invariably refer to "tail areas" of the density function of the distribution of the statistic. If 95% refers to the lower tail of the distribution of chi-squared then the Mah et al model should be rejected because the fit is too poor (upper tail area < 5%). If 95% refers to the upper tail of the distribution then the model should, perhaps, be rejected because the fit is too good (lower tail area < 2.5%) - for the observations to represent a random sample drawn from the population described by the model.) And Travis and Tucker do state (quite correctly) that "Isoeffect formulae will be useful only if there is some agreement between various centers on the fit of the model to their data." However, they offer neither measures nor criteria of "fit" for their ESD model of the Mah et al clinical data nor do they remark the absence of such measures and criteria in the two preceding studies. b) Point estimates of the parameters of the respective dose response model; c) Interval estimates (e.g., confidence limits) of the model parameters. Although one of the criteria given by Wara et al for preferring the ED over the NSD rival model of dose-response is that, "... the curves for ED appear steeper ..." and Mah et al and Travis and Tucker also discuss and compare the "steepness" and "slope" of the respective dose-response curves (ED and ESD), since sample estimates of the respective parameter vectors and their covariance matrices are not given in their paper, one cannot determine whether the apparent differences in the slopes are real or artifactual - a result of sampling variation. Thus, the absence of any of the required measures of uncertainty of their findings is further evidence of that lack of any very sophisticated stochastic world-view that we have remarked in virtually every one of the other studies in the field of radiation biology that we

have reviewed. As we have noted, it is this evident lack of sense of the contingent, or aleatory, element that is invariably present in all natural processes which leads to the presentation of all findings with the apodictic assurance that is the hall-mark of what Fleck has called exoteric science.

2) Figure 39a is a reproduction of Fig. 5 of the Mah et al study which is described as follows in their report: "Fig. 5. Dose-response curve for the incidence of acute radiation-induced pulmonary damage caused by fractionated radiation therapy in humans. The solid curve is the best fit sigmoid to the data points as determined by probit regression. An ED50 of 1000 ED units with a standard error of 40 (horizontal error bar on the curve) is predicted. Vertical error bars are the binomial standard deviation of the points." The Wara et al and Travis and Tucker papers present similar figures which are no more informative. Although the level of response - acute pneumonitis - of greatest clinical interest lies at, say 5% or 10% - or at most, 30%, each of the three studies only provides interval estimates of the respective 0.50 quantiles, that is, the "dose" that educes a 50% response (e.g., the ED50). And although the Wara et al and Mah et al reports give estimates of the standard deviations of each of the observed binomial responses, this information is irrelevant to the usefulness of the model; the information of clinical interest` is, of course, the size of the 0.95 (Or 0.99) confidence limits on the sample estimates of the lower quantiles, e.g., ED(0.05), ED(0.10), etc. (Wara et al present (1-α) interval estimates of the ED50 for which α is unspecified on the plots of their several dose-response curves. Mah et al present the 0.68 confidence interval on the ED50 (ED50 ± one standard error). Travis and Tucker present the 0.95 confidence limits on the ESD50).

And it must be remarked that, since the uncertainty in the sample estimates of the quantiles of a dose-response curve is a minimum in the vicinity of $\pi = 0.50$ (more precisely, it is a minimum in the vicinity of the centroid of the sample), it is rather unhelpful to provide interval estimates of the ED50 when the region of clinical interest lies well below that level of dose, i.e., at the ED(0.05), since the width of the confidence intervals increases quadratically with the "distance" from the centroid of the sample.

3) We also find in each of the three studies that, although the clinical data consist of the records of dose, fractions, time, and (binary) response of several dozen individual patients (Wara et al, n=42, Mah et al, n=54, Travis and Tucker, n=54) the data are arbitrarily aggregated - "binned" - according to the respective equivalent single "doses" - ED, ED and ESD - into only a few ($3 \leq n \leq 6$) dose groups. Neither the motivation nor the criteria for "binning" the disaggregated data are given in any of the three studies (It should be noted that several algorithms for optimal "binning" of multivariate data exist in the statistical literature). Moreover, the groups within each study are not uniform in either range of "doses" or numbers at risk included in each group. It is also the case that even for such coarse aggregations the numbers at risk at each level of"dose" are much too small: For Wara et al, the numbers at risk are $n_i = 34, 3, 4, 1$. For Mah et al, the numbers at risk are $n_i = 6, 12, 12, 10, 14$. (Note that $\Sigma n_i = n$). As remarked in earlier sections of this report, the numbers at risk at each level of dose should be $25 \leq n_i \leq 50$. See Finney, 1971. Moreover, for the Mah et al data the response in the lowest dose-group exceeds $= 0.30$ which means that the estimates of the quantiles of greatest clinical interest, ED(0.05) and ED(0.10), are extrapolations - and hence encumbered with an uncertainty of estimate, as described say by the 0.95 confidence limits, that increases quadratically with the difference in dose from the lowest observed level. Travis and Tucker do not give the numbers at risk for their six dose groups; however, we note (from their Fig. 2) that these groups include two extreme responses: one at 0% and one at 100%. Obviously, if the data were to be "binned" prior to the construction of a model then a sensitivity analysis should have been performed in which the effects of several grouping criteria (dose, numbers at risk, etc.) on both the point and set, or interval, parameter estimates of the model were assessed. Moreover, since the number of responders, r_i, in each group is so small (Wara et al, $r_i = 0, 1, 2, 5$; Mah et al, ri = 2, 5, 8, 9, 12) only a small change in the (arbitrary) limits of a "dose"-group can produce a nontrivial change in the observed response (e.g., for Mah et al, changing the upper limit on the lowest group from 949 to 952 changes the observed response

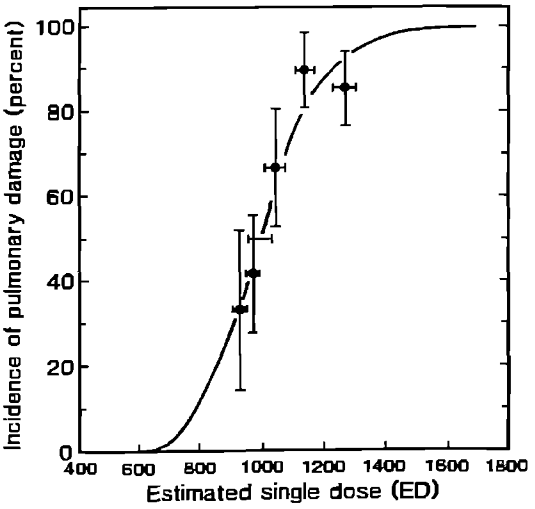


Fig. 39a. The figure is reproduced (with permission) from the report of Mah et al (1987): "Fig. 5. Dose-response curve for the incidence of acute radiation-induced pulmonary damage caused by fractionated radiation therapy in humans. The solid curve is the best fit sigmoid to the data points as determined by probit regression. An ED50 of 100 ED units with a standard error of 40 (horizontal error bar on the curve) is predicted. Vertical error bars are the binomial standard deviation of the points."

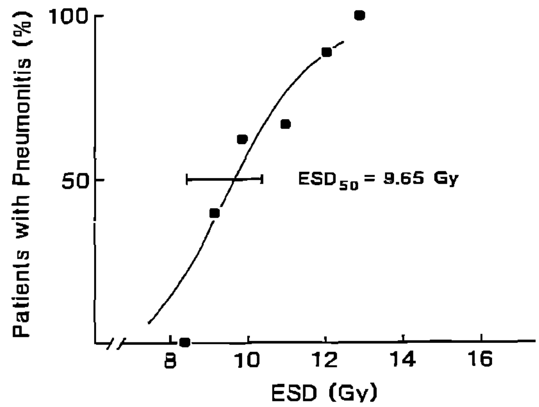


Fig. 39b. Reproduced from Fig. 2 of the Travis and Tucker 1987 report: "Incidence of pneumonitis in the data of Mah et al plotted as a function of calculated ESD ('Equivalent Single Dose'). The ESD is determined from the fractionation schedule using the linear quadratic model adjusted for the influence of treatment time." (Reprinted with permission from E. Travis and S. Tucker (1987).

from 2/6 = 0.33 to 3/7 = 0.43, a change of 30%).

It has been shown previously that dis-aggregated data should be binned, or grouped, post-model construction, on the criterion of the model-estimated probabilities, or risk, say $\hat{\pi}_i$, in order to assess the goodness-of-fit. These estimated probabilities are given by

$$\hat{\pi}_i = \Phi(\underline{x}_i^T \hat{\underline{\beta}})$$

for a probit model, where $\Phi(\ )$ is the distribution function of the standard Normal deviate and

$$\hat{\pi}_i = \exp(\underline{x}_i^T \hat{\underline{\beta}})/[1 + \exp(\underline{x}_i^T)]$$

for a logit model. Such aggregation into "deciles of risk" is necessary to improve the validity of the asymptotic chi-squared goodness-of-fit statistic. See Finney (1970); Hosmer and Lemeshow (1980); Hosmer and Lemeshow (1982; 1989). N.B. We have previously remarked in the literature several instances of similarly curious procedures: Just as "binning" disaggregated binary response data is appropriate if done post-hoc on the basis of the estimated probability π_i but not (pre-hoc) on the ED, so also the degree of "straightness" of a data plot is a useful criterion (of the concordance of model and data) if it is a probability plot but not an F_e-plot, and a geometric distribution of dose is appropriate for estimation of the parameter of a model of binary response on log dose but not on the dose itself. These and other similar instances of homeopathic magical thinking that we have noted are suggestive of the habits of thought of the South Pacific cargo cults.[17]

4) For the Travis and Tucker LQ model of clinical acute pneumonitis, it is evident from their Fig. 2 that at least half of the data are at levels of single dose > 10 Gy although the stipulated upper limit of validity of the LQ model is at 10 Gy (Fowler, 1984). Indeed, as has been recently remarked, "... LQ is not intended for doses higher than 8-10 Gy [per fraction] the highest used in radiotherapy" (Fowler, 1989). [It was found in secondary analysis that for 31 out of 54 patients (57%) the equivalent single dose (ESD) for the "LQ + time" model exceeded 10 Gy. The maximum value of ESD was 13.1 Gy.]

5) Although the clinical data in each of the three studies are multivariate, that is, the data include values for dose (D), fractions (N), and elapsed time of treatment (T) for each patient, each of the sets of investigators prefers to project, or (better) to map, these data into an equivalent single dose (ED or ESD) at N=T=1. But as this treatment regimen lies well out side the main body of observations, either of the two transformations represents a gross extrapolation.

This can be seen more clearly in Fig. 40. Figure 40a is a 3-dimensional scattergram of the Mah et al clinical data that were used in the Mah et al, and Travis and Tucker studies. The filled symbols identify those patients in which acute pneumonitis occurred; the open symbols identify those patients in whom it did not. Figure 40b presents the Box plots of the two conditional distributions of dose. Three of the weaknesses of this clinical data for construction of any model of dose-response are at once apparent in Fig. 40a. i) the presence of strong multicollinearity in the distribution of the covariates dose, fraction, and days; ii) the presence of (at least) two outlying and/or influential observations; iii) the presence of a high degree of overlap of the two conditional distributions of the covariates on the response. The weakness is shown quite vividly in the Box plot of Fig. 40b. These are, of course, the weaknesses that are commonly found in clinical data. See, for example, Figs. 4b and 36a for instances of multicollinearity and Fig. 2 of Appendix II for an instance of overlap and outliers.

Figure 41a shows the super-position of the scattergram of the equivalent doses (ED) for each patient in the Mah et al sample - the "cylinder" at N=T=1 - on the scattergram of Fig. 40a. Figure 41b is the projection of Fig. 41a onto the dose-fraction plane. Figure 41c is the projection of Fig. 41a onto the fraction-days plane. The ellipses in Figs. 41b and 41c are the 0.95 contour ellipses for the sample of data, size n=54.

It is, of course, well-known that the intercept of a regression model of a sample has no substantive meaning if the data do not include observations in the immediate vicinity of the origin. It is obvious from Figs. 41a-41c that the data of Mah et al do not include the observations that are required for the equivalent single doses (ED and ESD) to have substantive meaning.

The mapping of the set of clinical responses observed for N > 1 into the hypothetical response at N=1 that is accomplished by the Mah et al construction of the ED from their clinical

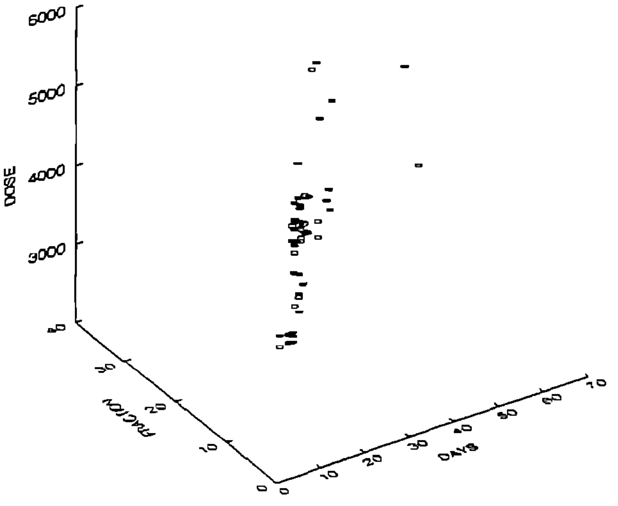


Fig. 40a. A three-dimensional scattergram of the n=54 treatment regimens of Mah et al. The 36 responders (radiation pneumonitis) are identified by filled circles; the 18 non-responders by open circles. It appears that the three variables are strongly correlated in these data. It also is evident that the two conditional distributions of the treatment variables (on response and non-response) overlap quite strongly. There appear to be two outlying observations (one responder, one non-responder).

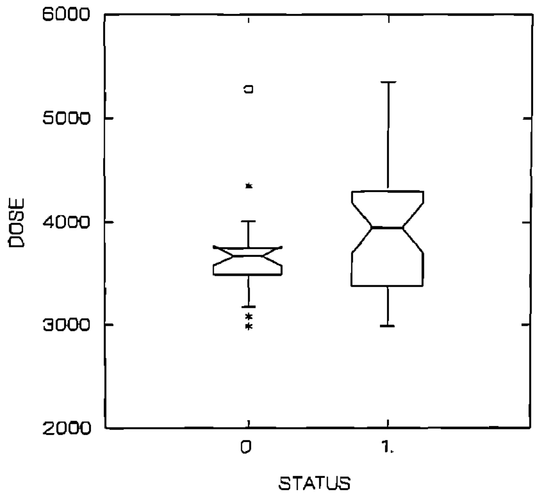


Fig. 40b. Box plots of the conditional distribution of dose. In the Box plot the horizontal line at the narrowest part of the central polygon represents the median (the 0.50 quantile) of the distribution. The lower and upper limits of the polygon represent the 0.25 and 0.75 quantiles, respectively. The upper and lower limits of the notch represent the 0.95 confidence limits on the median. The vertical lines include all of those observations within 1.5* interquartile range from the median. Observations that lie between 1.5 and 3 times the interquartile range from the median are identified by an asterisk; observations that lie beyond 3 times the interquartile range are identified by a circle. 0 denotes non-response; 1 denotes response. The overlap is obviously quite large as is suggested in Fig. 40a. The cognate plots of fractions and days disclose similar degrees of overlap of the conditional distributions. The presence of these overlaps of such a degree in the conditional distributions of treatment variables is one of the principal weaknesses of clinical data for estimation of useful dose-response models. See Appendix I.

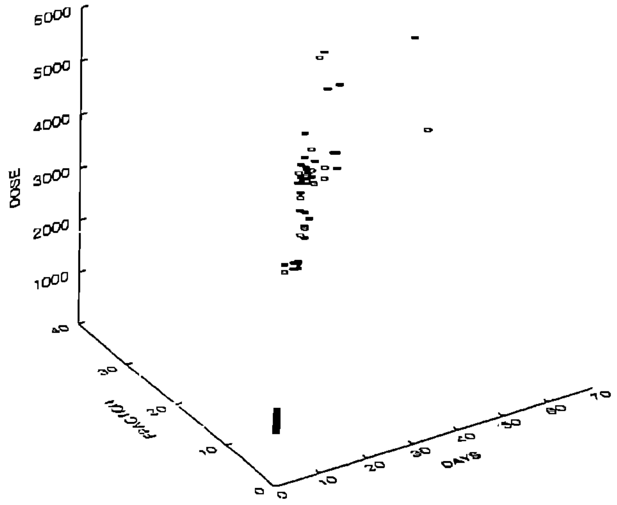


Fig. 41a. Superposition of the 3-dimensional scattergram of the Mah et al data of Fig. 40a and the 1-dimensional scattergram (at N=T=1) of the cognate values of the ED of the so-called Wara model. The figure describes the mapping of the sample observations into the equivalent single dose, ED of Wara et al:

$$D \xrightarrow[T^{0.058}]{N^{-0.377}} ED$$

It provides a vivid illustration of Kuhn's description of normal science as, "... a strenuous and devoted attempt to force nature into the conceptual boxes provided by professional education." Since the two scattergrams do not overlap the mapping describes an extrapolation of the dose-response information in the clinical data.

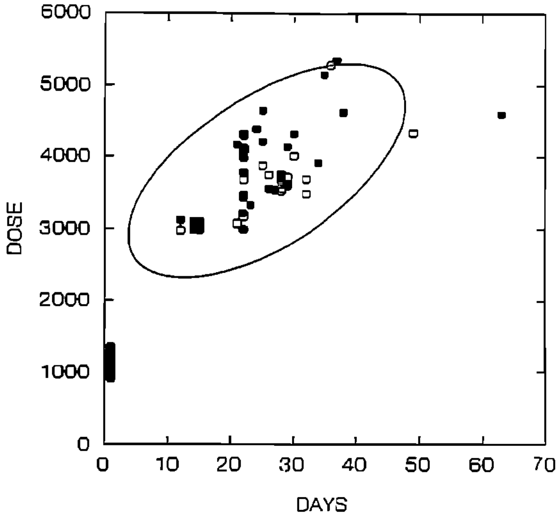


Fig. 41b. The projection of the two scattergrams of Fig. 41a onto the dose-days plane. The closed curve is the 0.95 contour ellipse on the distribution of the Mah et al data.

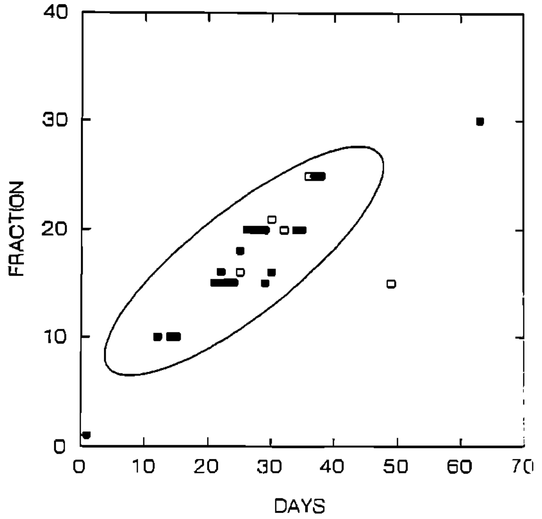


Fig. 41c. The projection of the two scattergrams of Fig. 41a onto the fraction-days plane. The closed curve is the 0.95 contour ellipse on the distribution of the Mah et al data.

The high degree of overlap of the two conditional distributions of the treatment regimens (D_i, N_i, T_i) in the Mah et al data is quite evident in Figs. 41b and 41c.

data and the Wara et al exponents and is described quite vividly in Figs. 41a-41c can be criticized on biological as well as statistical criteria: We have remarked previously that the response at N=1 differs qualitatively from that at N > 1 since in the former case the target tissues have not been previously irradiated and thus the responses of the two regimens may differ both in level and in kind; they may be incommensurable. In the physical metaphor proposed in section above, the Mah et al procedure is tantamount to replacing the steady state behaviour of an oscillator with the transient behaviour.

6) Although each study incorporates a priori information from animal experiments into the point estimates of the parameter vector of the respective models of clinical data, none of the three papers makes a very good case for doing so. Although it is obvious to any reader that the clinical sample data must be very weak - for the reasons remarked above - and hence would easily justify an attempt to use more robust non-sample information to "strengthen" (stabilize) the parameter estimates, this (central) issue in the modelling of clinical data is only obliquely referred to (if at all) as the motivation for the respective inter-species extrapolations of dose-response functions that are implicit in the transformations $D \longrightarrow ED$, and $D \longrightarrow ESD$.

7) In Wara et al the dose metameter is logED but in Mah et al, and Travis and Tucker it is ED and ESD, respectively, although both papers refer to the "Wara model". Since it is well-known, both empirically and from the Weber-Fechner "law", that for most toxic end-points, the level of a binary response is proportional to the level of the logarithm of the "dose" of the toxic agent - not to the dose itself - it would seem likely that the dose-metameter that implements the so-called "Wara model" is misspecified in both of the latter papers. [Subsequent secondary analysis will disclose this to be, in fact, the case.]

8) It appears from the Travis and Tucker paper that these two authors have misspecified the form of the response in their "LQ + time" model of the Mah et al data. Travis and Tucker describe their model as $E = \alpha D + \beta D^2/N - \gamma T$ where, "E = constant effect level". But, unless the response variable, E, is identified explicitly and specified precisely the model lacks any empirical content whatever. However, since their model is deployed to describe the binary response data of Mah et al, then E must be either a proportion, say π, or a function thereof, say $g(\pi)$. That is, either $E = \pi$, or $E = \Phi^{-1}(\pi)$, or $E = \log[\pi/(1-\pi)]$, or $2\sinh^{-1}\sqrt{\pi}$, or $E = \log(-\log\pi)$, i.e., the probit, logit, inverse hyperbolic sine, and extreme value transforms, respectively. But, each of these transforms requires that a constant term, say β_0, be included in the specification of the linear predictor, $\eta = \underline{x}^T\underline{\beta}$. Thus, it appears from the context of their equation that $E = \pi$ for the "constant effect level" of the Travis and Tucker model. But $E = \pi$ implies, of course, that the tolerance distribution is uniform (see Figs. 31b and 31c of this report) which is, a priori, quite impossible.

Moreover, although the observed response, π_i, $1 \leq i \leq n = 54$, lies in the interval, $0 \leq \pi_i \leq 1$, it may well be that for some values of the treatment variables, say $\underline{x}_i^T$, the estimated response, $\hat{\pi}_i = \underline{x}_i^T\underline{\hat{\beta}}$, lies outside this interval, $\hat{\pi}_i < 0$ or $\hat{\pi}_i > 1$, and hence neither the estimate nor the model by which it was obtained has any biological meaning. [In our secondary analysis of the Travis and Tucker LQ + time model, $\pi = \alpha D + \beta D^2/N - \gamma T$, this was found to be the case for three observations in the Mah et al data.] Therefore, it appears that for the Travis and Tucker version of the LQ + time model, the form of the distribution of the response is misspecified. The error is exactly the same as that committed by Shellabarger et al (1969) nearly 20 years earlier. See Annex III and Figs. 13 and 30-33.

9) In all three papers (Wara et al, Mah et al, Travis and Tucker) the a priori information on the parameter vector of the model of the clinical pneumonitis data that is obtained from the mouse pneumonitis experiments is deployed without regard to either sampling errors of parameter estimate within the respective clinical and animal studies or to the errors of relevance that encumber the extrapolation of this information between the two species, i.e., the mouse-to-man, etc., errors. (These are the errors denoted as ϵ_{ij} and δ_{ij} in the DuMouchel and Harris (1983) study described above in section 14.) The argument given by Mah et al for combining animal and clinical data and for choosing the Wara et al estimates for the exponents of N and T quite ignore

the ingenuous disclaimer which is offered by Wara et al ("Naturally the confidence in this slope is limited since its line was derived from only two points.") for their estimates, 0.377 and 0.058, of the parameters of their model, "Total dose = $EDN^{0.377}T^{0.058}$." This remark should have been a sufficiently startling admission to persuade most investigators either to independently validate these estimates, or to seek other estimates of these exponents. [But we recall that the von Essen model of volume effects also has been widely adopted despite the explicit disclaimer in his 1960 paper warning that his estimates of the volume exponents in his model are based upon fictitious ("hypothetical") isoeffect curves. See Appendix I. And as we have remarked above in section 2.3, it often appears that many scientists do not read - or else read without result - the reports of previous studies that they cite.] Moreover, in each of the three studies the respective models of the animal and human data are not isomorphic: For Mah et al it is isoeffect vs probit dose-response. For Travis and Tucker it is isoeffect vs linear and logit dose-response. [For this reason, even though both Mah et al and Travis and Tucker assume that the parameter information obtained from the mouse data has infinite precision - no standard errors of estimate for either the exponents of N and T (Mah et al) or the ratios α/β and γ/β (Travis and Tucker) are given in the respective reports - the precision of the posterior estimates, $\hat{\beta}$, of the univariate probit (logit) models on the respective equivalent single doses (ED and ESD) was found in our secondary analysis to be too small to be of any use in either scientific explanation or clinical exploitation of the dose-response relationship described by the respective models: [$\hat{\beta}_j^{**}/\sqrt{Var(\hat{\beta}_{j**})} \sim 2.9$].

10) It appears from close examination of Fig. 1 of the Travis and Tucker paper, that the plot of $D(\alpha/\beta + D/N)$ vs T, from which they deduce the linearity of the term, $f(T) = \gamma T$, representing the effect of time in the LQ + time model, that the plots of both sets of data (Wara et al and Field et al, 1976) may be described as "concave down" rather than linear - if the remote observation at T = 52 days in the data of Fields et al, 1976 is deleted. This suggests an alternative parameterization of the time factor: $f(T) = \delta \log T$. [Actually, of course, the statistically correct procedure for determining whether any given model underfits a set of data - the matter that was at issue in their study - must be based on a plot of the residuals of the current model, that is, the set of differences of observed and estimated responses, against the added variable; in the Travis and Tucker study this variable was the elapsed time T.]

16.5 Secondary analyses of the Mah et al, and Travis and Tucker studies.

As noted above, the Mah et al study does not provide a statistically adequate description of the so-called "Wara model" of their clinical data. Therefore, we present in Table 5a the results of a secondary analysis of the Mah et al data based upon their Table 5 from which the dose-response curve in their Fig. 5 (our Fig. 39a) was constructed. We note that the Wara model does indeed fit their data - on the evidence of an aggregate statistic, the Pearson chi-squared. However, the precision of the parameter estimates as measured by the ratio (Student's) $t_j = \bar{\beta}_j^{**}/\sqrt{Var(\hat{\beta}_j)}^{**}$, is rather small, $t_j - 2.50$, suggesting no more than that the relationship may be "real", that is, not an artifact of sampling; it is not sufficiently well-defined to be very useful. Note that for a useful model the statistic $g(=t^2 Var(\hat{\beta}_1)/\hat{\beta}_1^2)$ should be less than 0.05 for t = 1.96. (Indeed, as will become apparent below, the dose-response relationship described by Mah et al yield estimates of the lower quantiles, say ED(0.05) that are sufficiently imprecise to be rather "dangerous to your health", if it were to be deployed in the clinic.) Moreover, as we remarked above, the aggregation of the 54 observations into n=5 dose groups or "bins" is quite arbitrary. The preferred procedure for such data, and one which we have followed throughout this report, is to construct the model directly from the unaggregated observations, i.e., n=54. See Finney 1973 and Hosmer and Lemeshow, 1989. The cognate ED model of the unaggregated data is described in Table 5b.

A better warrant for the clinical usefulness of any predictive model of normal tissue response is described by the interval estimates of the smaller quantiles of the dose-response curve, i.e., ED(0.10) or ED(0.05) rather than by those for the ED(0.50) that were given in Mah et al. [N.B. "The size of the confidence limits is inversely proportional to the quality of the data used to make the estimate and directly proportional to the amount of extrapolation involved. This

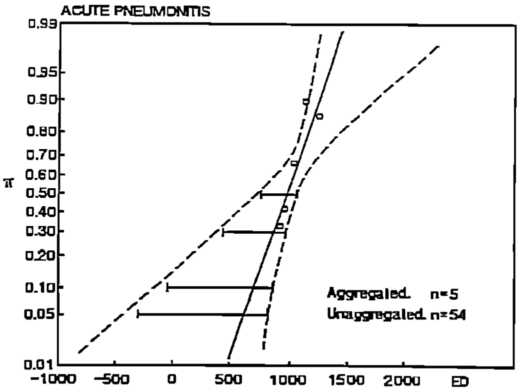


Fig. 42a. Plot of the dose-response curve and the 0.95 confidence limits thereon for two probit models, $z = \beta_0 + \beta_1 ED$, of the Mah et al data. The scattergram of the n=5 groups of aggregated observations constructed by Mah et al is super-posed. The solid line is the dose-response curve for the probit model of the unaggregated (n=54) data. The dashed lines describe the 0.95 confidence limits thereon. The horizontal lines describe the 0.95 quantiles for the probit model of the aggregated (n=5) data.

It is quite evident that the Mah et al probit model misspecifies the dose metameter since the lower limb of the 0.95 confidence limits on the quantiles of ED are negative for the lower quantiles which are of greatest clinical interest, e.g., ED(0.05), ED(0.10). The figure suggests that the appropriate dose metameter is logED. Subsequent analysis confirms this suggestion: The model of choice is $z = \beta_0 + \beta_1 \log ED$. See Table 5c.

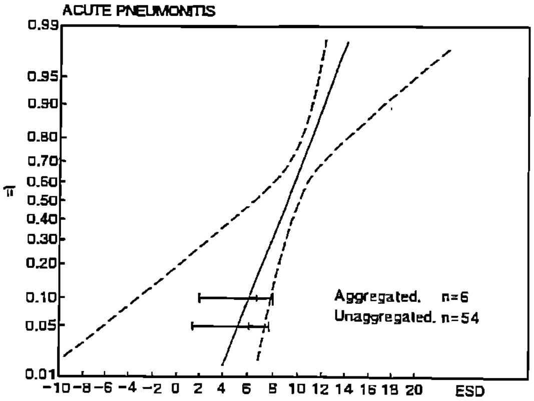


Fig. 42b. Plot of the dose-response curve and the 0.95 confidence limits thereon for two probit models, $z = \beta_0 + \beta_1 ESD$, of the Mah et al data. The solid line is the dose-response curve for the probit model of the unaggregated (n=54) data. The dashed lines describe the 0.95 confidence limits thereon. The horizontal lines are the 0.95 confidence limits on the 0.05 and 0.10 quantiles for the logESD, showing that the logarithmic transformation corrects the misspecification error of the dose metameter that is present in the Travis and Tucker report. The model of choice is $z = \beta_0 + \beta_1 \log ESD$.

Note that Figs. 42a and 42b provide vivid evidence of the incorrect choice of dose metameters in both the Mah et al and Travis and Tucker studies since the lower limb of the 0.95 CL is negative for the lower quantiles of both ED and ESD.

Table 5a. Probit Model of Mah et al (1987) Clinical Data based on Wara ED.

$z = -4.716 + 4.777$ ED. n = 5
(-2.633)# (2.855)

Pearson chi-squared, RSS = 1.879 on 3 df. $P(\chi^2 > RSS|3) = 0.402$. g = 0.471##

$t_j = \bar{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$. ## $g = t^2 Var(\hat{\beta}_1)/\hat{\beta}_1^2$ for t = 1.96

response	quantile	lower 0.95 CL	upper 0.95 CL
0.50	987.28	772.49	1064.37
0.05	643.96	-287.39	825.35

See Fig. 42a.

Table 5b. Probit Model of Mah et al (1987) Clinical Data Based on Wara ED.

$z = -4.481 + 4.562$ED. n = 54
(-2.559)# (2.788)

Pearson chi-squared, RSS = 55.21. df = 52.

$P(\chi^2 > RSS|52) = 0.64$. g = 0.494##.

$t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta})}$. ##$g = t^2 Var(\hat{\beta}_1)/\hat{\beta}_1^2$ for t = 1.96

response	quantile	lower 0.95 CL	upper 0.95 CL
0.50	982.30	741.83	1062.84
0.05	622.78	-431.76	815.35

See Fig. 42a.

Table 5c. Probit Model of Mah et al (1987) Clinical Data Based on logED.

$z = -36.789 + 12.284$logED. n = 5
(-2.886)# (2.915)

Pearson chi-squared. RSS = 1.596 on 3 df.

$P(\chi^2 > RSS|3) = 0.340$. g = 0.452##

#$t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$. ##$g = t^2 Var(\bar{\beta}_1)/\hat{\beta}_1^2$ for t = 1.96

response	quantile	lower 0.95 CL	upper 0.95 CL
0.50	987.87	821.86	1058.69
0.05	726.49	332.10	853.32

important information is lost if the confidence limits and best estimates are not routinely reported. The width of the confidence interval is one of the best measures risk assessors, and risk managers have to evaluate the quality of the estimates of potential risks. It is important to distinguish between those situations in which the risk is precisely estimated and those in which it is not" (C. Park and R. Snee, 1983)] We give the 0.95 confidence limits on the ED(0.05) quantile for the Wara model of the Mah et al data in Table 5a for both aggregated (Table 5a) and unaggregated (Table 5b) data. These estimates would seem to almost completely discredit the Mah et al explanation of their data by the ED model: The lower 0.95 confidence limits on the quantiles of clinical interest, ED(0.50), ED(0.10), etc., are negative! But, an important principle of model construction is to "... choose a model that is consistent with the data and yields parameter estimates consistent with ... prior beliefs (J. Robin and S. Greenland, 1986). A priori, ED(π) < 0, $0 \leq \pi, \leq 1$, is most unlikely.

The probit dose-response curve for the Wara model of the Mah et al data and 0.95 confidence limits are presented in Fig. 42a. It is obvious from this figure that, contrary to the assertions of Mah et al and Travis and Tucker, the Mah et al clinical pneumonitis data do not yield "a ... well-defined dose-response curve", (Travis and Tucker 1987) nor yet even "a distinct dose-response curve for human pulmonary tissue to fractionated radiotherapy" (Mah et al, 1987). The ambiguity with which the lower quantiles of the curve are defined by their model almost surely enshrouds the possibility of a clinical misadventure or two, if the model were deployed. (N.B. It will be recalled that the lung is a Class I organ "... in which radiation lesions [acute and chronic pneumonitis] are fatal or result in severe morbidity." Radiation Biology and Radiation Pathology Syllabus. R.T.1. 1975.)

Figure 42b is the probit dose-response curve for the Travis and Tucker equivalent single dose (ESD) model of the Mah et al data. The Travis and Tucker model of These data is reproduced in Fig. 39b. It is seen at once that it exhibits the same ontological weakness as does the Wara model, namely, the lower limits of the 0.95 confidence interval for the quantiles of clinical interest. ESD(0.10), ESD(0.05) are negative. It exhibits the further conceptual weakness in that the majority of the values of ESD lie beyond 10 Gy, the stipulated upper limit of the validity of the LQ model (Fowler, 1984, 1989); indeed several lie well-beyond the well-known 10 Gy limit.

It is obvious from Table 4a and Fig. 42a that, as was anticipated above, the dose-metameter - ED - is misspecified in the Mah et al version, since the lower limb of the 0.95 confidence limits are negative for the lowest quantiles. This suggests the alternative parameterization of the model, $z = \beta_0 + \beta_1 \log ED$. It is shown in Table 5c that this logarithmic transformation of the ED rectified the ontological weakness - misspecification of the dose-metameter - of the Mah et al model. A similar transformation of the ESD, to give the model $z = \beta_0 + \beta_1 \log ESD$, is also required. See Fig. 42b. Note that although the lower limb of the 0.95 confidence limits on the 0.05 quantiles is now non-negative the uncertainty with which this clinically important dose can be estimated from these data is enormous: The width of the 0.95 confidence limit on ED(0.05) is 5.21 Gy!

Figure 43a presents the Box plots of the conditional distributions of ED for the Mah et al data and Fig. 43b is the cognate plot for the conditional distributions of ESD. Comparison with Fig. 40b shows that the overlap of the conditional distribution of "dose" has been reduced by the two transformations D ⟶ ED and D ⟶ ESD although it is still quite large. Figure 43c is a plot of ESD vs ED.

The correct multivariate procedure to deploy to minimize the overlap of two conditional distributions is discriminant analysis. This procedure will provide estimates of the parameter vector $\underline{\beta}$ of a logit dose-response model for unaggregated data such as that of Mah et al. See Appendix I. However, unless the conditional distributions are multivariate Normally distributed these estimates may be biased although such estimates now appear to be more robust than was thought to be the case earlier. Moreover, if the two conditional distributions are multivariate Normal, then the discriminant analysis estimates of $\underline{\beta}$ are more efficient than the estimates obtained by rival procedures such as logistic regression analysis. However, we have elected to obtain unbiased maximum likelihood estimates of $\underline{\beta}$ by the rival logistic regression procedures as described in Table 6c below.

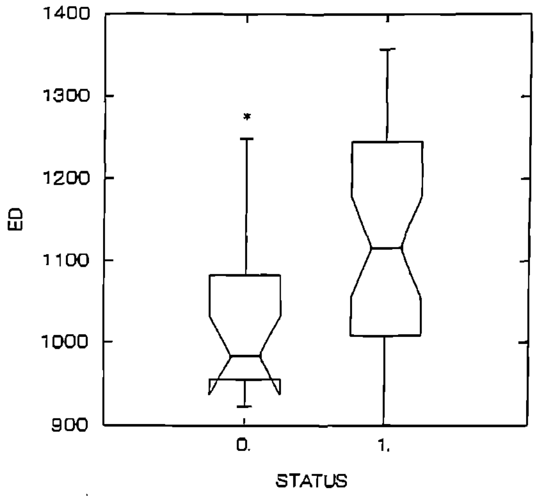


Fig. 43a. The figure presents a superposition of the Box plots of the conditional distributions of ED on response, or status, for the Mah et al data. Status = 0 (non-response); Status = 1 (acute pneumonitis). The transformation D ⟶ ED has reduced the overlap of the conditional distributions of dose described in Figs. 40a and 40b.

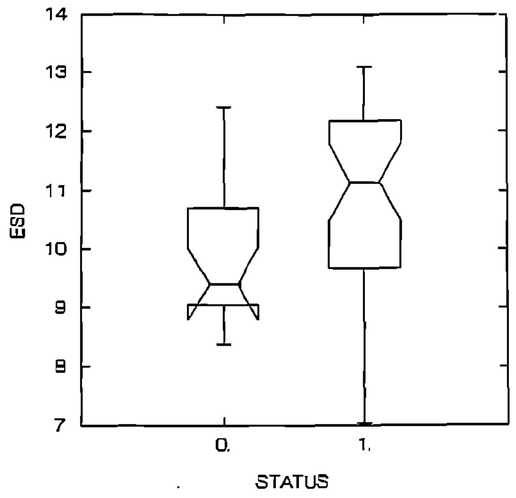


Fig. 43b. The figure presents a superposition of the Box plots of the conditional distributions of ESD on response, or status, for the Mah et al data. Status = 0 (non-response); Status = 1 (acute pneumonitis). The transformation D ⟶ ESD has reduced the overlap of the conditional distributions of dose described in Figs. 40a and 40b.

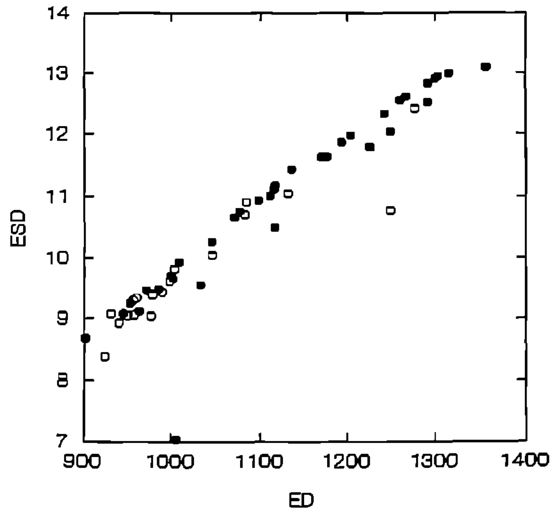


Fig. 43c. Plot of the ESD vs ED for the Mah et al data. The filled circles identify the responders (radiation pneumonitis); the open circles the non-responders. The two outlying observations of Fig. 40 are clearly evident. Note the high degree of overlap of the conditional distributions of ED and ESD. Note also that most of the levels of ESD exceed 10 Gy, the stipulated upper limit of the ESD.

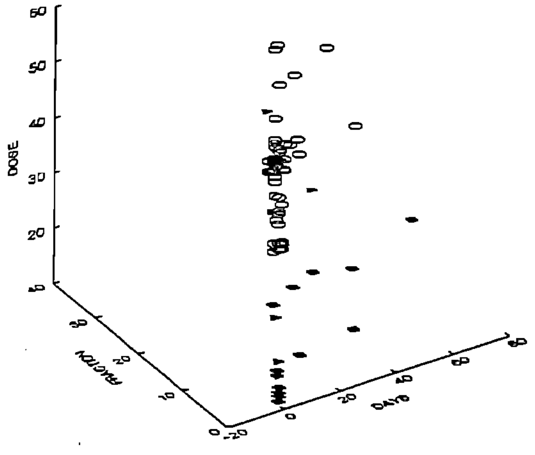


Fig. 44a. The figure presents a superposition of the three-dimensional scattergrams of the 54 clinical observations in the Mah et al study (circles), the 6 experimental observations on mice of the Wara et al study (triangles), and the 13 experimental observations on mice of the Field et al study (squares). The dose is the average corrected lung dose for each patient in the Mah et al study, the LD50/160 in the Wara et al study and the LD50/40-180 in the field et al study. The number of animals from which the LD50 estimates were determined is not given in either paper.

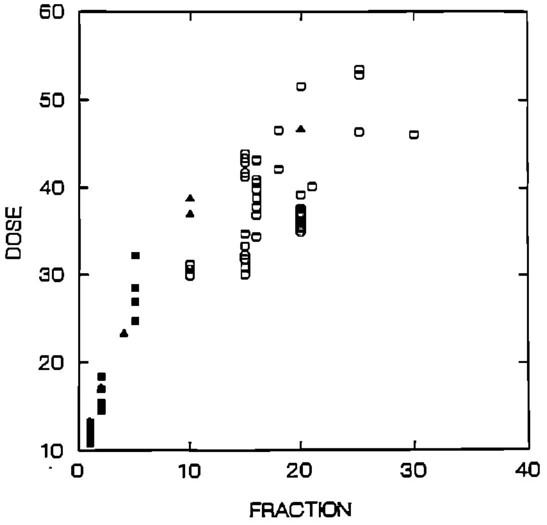


Fig. 44b is the projection of the observations in Fig. 6a onto the dose-fraction plane.

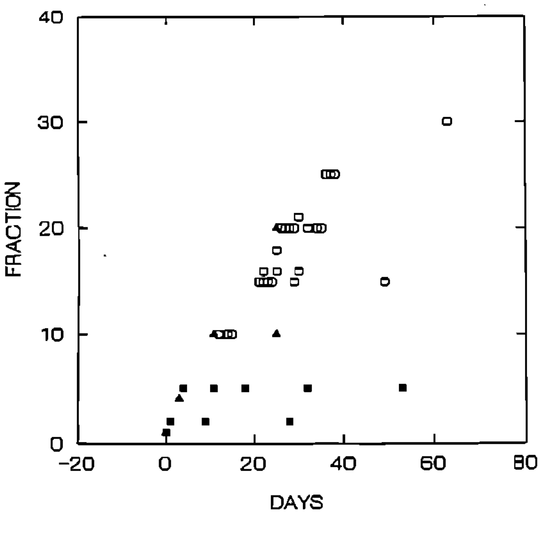


Fig. 44c is the projection onto the fraction-days plane. It is obvious from these three figures that there is very little overlap of the three conditional distributions. Hence, the deployment of the mouse information on dose-response to stabilize the parameter estimates of any dose-response model of the patient data require extrapolation between three regions of dose-time-fraction space as well as between two species; that is, a scaling problem encumbers the animal information.

Figure 44a is a super-position of the plots in the dose-fraction-days space of the respective data sets of the Wara et al (mouse), the Field et al (mouse), and the Mah et al (human) studies. Figure 44b is the projection of these data on the dose-fraction plane and Fig. 44c is the projection on the fraction-days plane. In all three figures the Wara et al data are identified by the triangles, the Field et al data by the squares, and the Mah et al data by the circles. For the Mah et al data each circle represents the average lung dose of low LET photons (Co^{60} - 25 Mev X-ray) in a single patient. For the Wara et al and Field et al data, each symbol represents the LD50 for radiation pneumonitis in mice (determined from groups of unspecified sizes) for 250 kvp X-rays (Field et al) and 300 kvp X-rays (Wara et al).

It will be noted at once from these figures that the distributions of the animal and patient data do not overlap in the space of the treatment variables. In fact, the distribution of the Mah et al patient data barely intersects the distribution of the mouse data of the Field et al study and the overlap of these patient data with the Wara et al mouse data is only somewhat greater. Thus, not only are the respective models of the mouse and patient data sets not isomorphic (i.e., the information in the mouse data is conveyed in isoeffect models, that in the patient data is summarized by a dose-response model), but the data sets themselves are also inconsistent in important features: the respective distributions do not overlap. Such animal isoeffect studies are intrinsically weak sources of relevant a priori information on the parameters of dose-response models of patient data since they are not in parallel with the patient studies and their data are distributed over a different region of the treatment variable space than are the patient data.

From Figs. 40 and 44 it appears that the weaknesses of both the sample and non-sample information for estimation of model parameters can be vividly summarized in terms of the several conditional distributions of the treatment variables, dose, fraction and time. 1) The weaknesses of the Mah et al clinical sample data for direct estimation of the parameter vector of any dose-response model resides in the large overlap of the two conditional distributions of dose, fractions, and time on the binary response. 2) The weakness of the Wara et al and Field et al animal non-sample information on $\underline{\beta}$ that can be obtained from the mouse data resides in the small overlap of the three conditional distributions of D,N,T on species.

Thus, the use of the animal data to construct the equivalent single doses of Wara et al (ED) and Travis and Tucker (ESD) requires extrapolations of information between two rather disparate sets of levels of (common) treatment variables as well as between two quite distinct species. That is, there is a scaling problem inherent in both procedures. Therefore, the two errors of relevance - in both species and in level and range covariates - that encumber the non-sample information on $\underline{\beta}$ and thus the parameter estimates of the ED and ESD models of the Mah et al data may be - indeed, almost certainly are - both quite large and largely unknown. Figure 45a is reproduced from Fig. 1 of the Travis and Tucker 1987 report: "Data of Wara et al (triangles) and Field et al (circles) for LD_{50} from pneumonitis is mice plotted as a function of the overall treatment time according to Eq. (3). The regression lines yield estimates of γ/β = 2.2 Gy^2/day for each data set." Figure 45b is the scattergram of Fig. 45a from which the observation at T=53 days has been deleted. The smoothed curves were obtained by the LOWESS procedure (Cleveland, 1979) with window width f=1.0 that is, 100% of the data were used to determine the location of each smoothed point. The trend of the data is obviously, concave down, rather than linear as reported by Travis and Tucker. Note that the respective radii of curvature are roughly equal. Figure 45c is a scattergram of the Field et al data of Fig. 45a. The curve is obtained by the LOWESS procedure with window width f=0.5. It is obvious that these data may be heterogeneous since the observation at T=53 days does not lie near the trend curve of the remaining observations.

The scattergrams of the Wara et al and Field et al data from which, in a kind of meta-analysis, Travis and Tucker obtained their estimates of the parametric form for their time factor (-γT) for the LQ model are presented in Figs. 45d and 45e, respectively. The ordinate is $D(\alpha/\beta + D/N)$ Gy^2 for α/β = 3 Gy for each plot. The a priori estimate, α/β = 3 Gy, was obtained by Travis and Tucker, as a consensus, from several previous reports, including Travis and Tucker, 1986. The Wara et al data of Fig. 45d are homogeneous or at least there are no evident

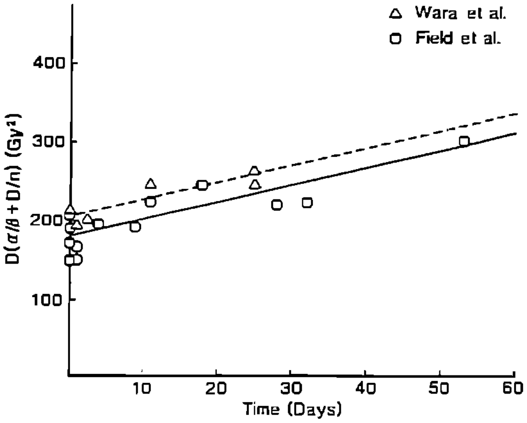


Figure 45a. Reproduced from Fig. 1 of the Travis and Tucker 1987 report: "Data of Wara et al (triangles) and Field et al (circles) for LD_{50} from pneumonitis in mice plotted as a function of overall treatment time according to Eq. (3). The regression lines yield estimates of γ/β = 2.2 Gy^2/day for each data set." The estimate of α/β = 3.0 Gy was chosen a priori by Travis and Tucker. (Reprinted with permission from the Intl. J. Radia. Oncol. Biol. Phys. 13: E. Travis and S. Tucker, "Isoeffect Models and Fractionated Radiation Therapy." 1987. Pergamon Press Ltd. Note that Fig. 45a represents yet another 'strenuous and devoted attempt" by radiobiologists to improperly stuff Nature into the conceptual "box" y = a + bx that is "... supplied by professional education" (T. Kuhn, 1970). We will expose four of the more glaring weaknesses of this particular attempt below: Fig. 45b, d, and e (The data in Fig. 45a are curved - "concave down" - not linear; Fig. 45c (The observation at T = 53 days does not belong with the remaining 12 observations in the Field et al study); Fig. 45f (The observation at T = 53 days dominates the estimate of the slope of the linear model of Fig. 45a); Figs. 45g and h (The appropriate geometric figure for a regression model of multivariate (D(x), N, T) data is a surface not a line).

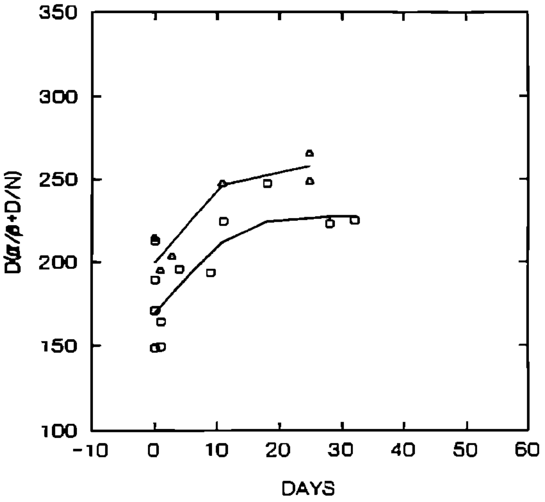


Figure 45b is a plot of the data of Fig. 45a, save for the observation at T = 53 days. The two curves are non-parametric regressions obtained by Cleveland's LOWESS procedure with the width of the smoothing window f = 1.0, that is, all of the observations in each of the two studies (Wara et al and Field et al) contribute to the smoothed estimates of their respective curves. Obviously, the linear model of Travis and Tucker does not capture the correct shape of the "time factor" for these data.

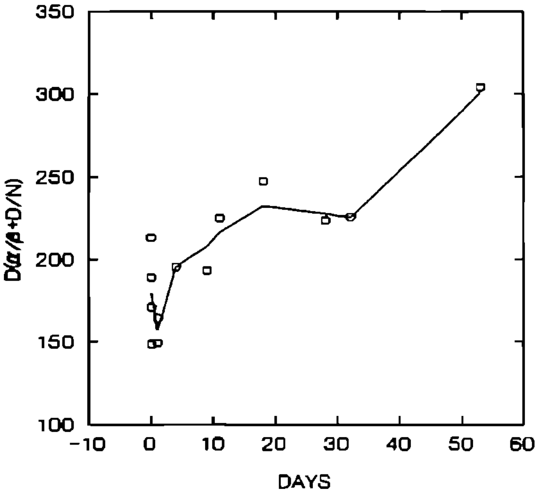


Figure 45c is a scatterplot of the data of the Field et al study of Fig. 45a. The curve is the non-parametric regression obtained by Cleveland's LOWESS procedure with f = 0.50. The shape strongly suggests that the observation at T = 53 days is markedly different from that of the remaining 12 observations.

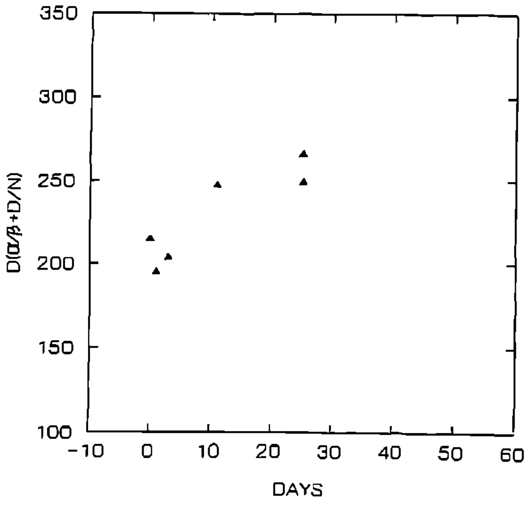


Fig. 45d is a scattergram of the data of the Wara et al study in the $D(\alpha\beta/ + D/N)$ - days plane for $\alpha/\beta = 3$ Gy. The plot is obviously concave-down rather than linear.

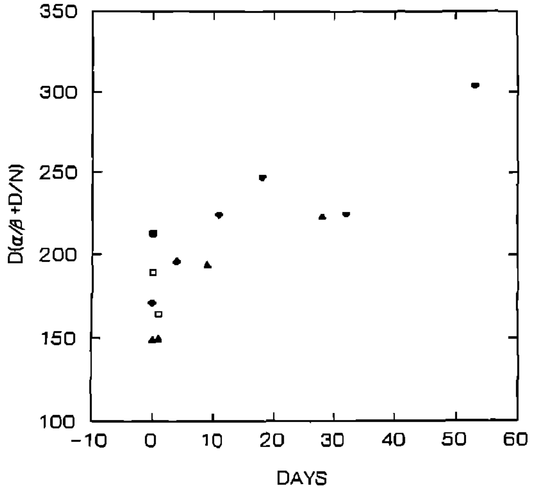


Fig. 45e is a scattergram of the sub-set of observations from the Field et al study that were selected by Travis and Tucker. Figures 45a and 45b describe a decomposition, by study, of the data presented in the Fig. 2 of the Travis and Tucker report. Figure 45b describes a further decomposition, by experiment, of the Field et al data selected by Travis and Tucker. The triangles represent the 4 observations from experiment LXI. The (open) squares represent the 2 observations from LXII. The (filled) circle represents the single observation from LXIII. The rhombi represent the 6 observations from experiment LXV.

Note that the respective plots of data of the Field et al experiments LXI and LXV and of the Wara et al experiment are all concave-down rather than linear. The most parsimonious representation of the true dependence of $D(\alpha/\beta + D/N)$ on days is $\delta\log T$ rather than γT.

It should be noted that the observation at 32 days in experiment LXV lies well below the trend line of the remaining five observations, suggesting that it is an outlier of those data.

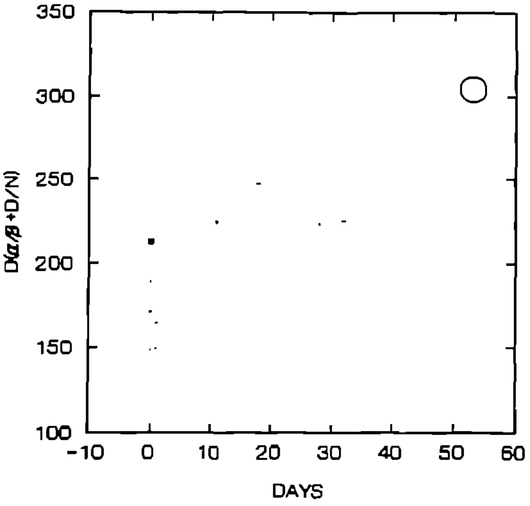


Figure 45f. Scattergram of the row-deletion diagnostic $R - R_{(i)}$ for $D(\alpha/\beta + D/N)$, T and $\alpha/\beta = 3$ Gy where R is the Pearson correlation coefficient for the subset of 13 observations from the Field, et al, 1977 data selected by Travis and Tucker (Compare with Fig. 45b). The size of the symbols is proportional to $R - R_{(i)}$ for the i^{th} observation. Filled symbols identify observations whose deletion increases the value of R. Open symbols identify observations whose deletion decreases the value of R. Obviously the linearity of the relationship between $D(\alpha/\beta + D/N)$ and T is dominated by the observation in experiment LXV at T = 53 days. The other influential observation is the single dose observation from experiment LX, the only observation from that experiment that was included in the Travis and Tucker selection.

features by which they may be further subdivided. However, this is not the case for the Field et al data of Fig. 45e. These data may be grouped according to the experiment by which they were obtained. Figure 45e is a super-position of the scattergrams of observations from the LXI, LXII, LXV, and LXIII experiments in the Field et al study: LXI consisted of observations on three treatment regimens of 2 fractions each in 1, 9, and 28 days plus a single-dose treatment regimen. LXII consisted of observations on one treatment regimen of 2 fractions in 1 day plus a single-dose treatment regimen. LXV consisted of five treatment regimens of 5 fractions each in 4, 11, 18, 32, and 53 days plus a single dose treatment regimen. LXIII consisted of one regimen of 8 fractions in 6 weeks and one of 30 fractions in 6 weeks plus a single dose regimen. However, Travis and Tucker, 1987 included only the single dose regimen from experiment LXIII of the Field et al study; the two observations at 6 weeks were omitted. (Note that in neither the Wara et al nor Field et al experiments is there strong evidence of a factorial design in the distribution of treatment variables from which to construct useful sample estimates of the coefficients of N and T in a multivariate model of dose-response - and hence valid inferences on the size and direction of their respective roles in a dose-response relationship.)

It is immediately evident from the Fig. 45e that there is considerable variation between the four experiments in the Field study; that is, there may be a large between-experiments effect. This conclusion is reinforced by comparing the respective differences between all 6 (= 4!/2!2!) pairs of values of LD50 for the four single-dose regimens with the average standard error of the difference between any two, we find the following set of ratios: 4.50, 1.56, 2.82, 2.94, 1.68, and 1.26. Although there was not enough information given in the Field et al report for us to assess the statistical significance of these differences, it is clear that three of the six are considerably larger than the average standard error of the difference. But, "Before combining results, however, we must consider whether the results of different studies are homogeneous." (Halvorsen, 1986). The experiments of Field et al are, obviously, not homogeneous.

There also appears to be considerable variation within the experiment LXV of the Field et al data: The observation at 32 days falls well below the curve described by the remaining five observations in this experiment; it could be considered an "outlier" with respect to the remaining observations within experiment LXV.

Note in Fig. 45e that the plots of the observations within experiments LXV (5 obs) and LXI (4 obs) are clearly concave down - as is also the plot of the 6 observations of Wara et al in Fig. 45d. Thus, the straight-line models that are fitted to these data in Travis and Tucker, 1987 do not accurately describe their data. Their Fig. 1 provides an instance of Simpson's paradox: The data within the individual experiments of Field et al describe quite a different relationship between $D(\alpha/\beta + D/N)$ and T than does the pooled data, i.e., curves (concave-down) vs linear. N.B. "One of the potential hazards of basing an estimate or test statistic on pooled data from several studies is known as Simpson's paradox. When this paradox arises, the conclusion reached in the meta-analysis may contradict the conclusions of the contributing studies." (Halvorsen, 1986). (N.B. It will be recalled that it was shown in Annex III that the empirical evidence for the curvilinearity - vs linearity - in the relation between gamma dose and leukemia incidence in the BEIR III 1980 report - that is, the LQ-L vs L-L models - also resides in a single anomalous observation. The similarity of several of the principal features of the data of the Field et al report to that of the Sparrow et al report (See Figs. 1a and 1b) should also be noted: For example, there is in both studies a high degree of heterogeneity in the pooled data of 4 experiments. Moreover, in both there is the presence of a version of Simpson's paradox in the inferences on the shape of the curves: It will be recalled that it was shown that evidence for the curvilinearity of the dose-response curve - LQ model - for radiation induced pink mutants in Tradescantia could be found in the pooled data, but in the data of only one of the four component experiments that were included in the "pool".

Note that the extreme observation in experiment LXV (at 53 days) in the Field et al study is a high leverage observation. This is shown more vividly in Fig. 45f which is a scattergram of the diagnostic R-R(i) of the Field et al data that were selected by Travis and Tucker (See Fig.

45b). In Fig. 45f the relative size of each symbol describes the effect on the Pearson correlation coefficient, R, of the sample, of deletion of that observation. Observations whose deletion decreases R are identified by open symbols; and conversely. Note that most of the 13 observations have very little influence on R. Indeed, it can be seen from Fig. 6c that the only two observations to have any influence on the linearity of the relation between $D(\alpha/\beta + D/N)$ and T are the two observations of Fig. 45e which have other dubious features that we have already remarked. However, the degree of linearity in the relation between the two variates is dominated by the observation at T = 53 days (index number i=12): For n = 13 the adjusted multiple correlation coefficient is $R^2 = 0.715$; for n = 12 it is $R^2_{(12)} = 0.455$; R is deflated by a factor of 0.636. That this observation provides much of the empirical evidence for the linearity of the relation between $D(\alpha/\beta + D/N)$ and T that is offered by Travis and Tucker can be shown by comparing the respective p-values for a test of the hypothesis on the parameter, α_2, of an additional quadratic term, T^2, in the model, $LHS = \alpha_0 + \alpha_1 T + \alpha_2 T^2$. With the observation at T = 53 days included, (n = 13), the additional term is rejected: p = 0.965. If this observation is omitted (n = 12) the additional term is not rejected at p = 0.106; a plot of the relationship between $D(\alpha/\beta + D/N)$ and T is not linear but concave-down as we have previously noted. A more parsimonious representation of this feature - concave down - of the plot of Fig. 45e is achieved by the transformation T logT, i.e., the model $D(\alpha/\beta + D/N) = \delta_0 + \delta_1 \log T$.

Thus, it appears that the conclusion of the Travis and Tucker 1987 analysis of the Wara et al and Field et al mouse data based on the procedures that are described in their paper, namely, that the time factor for the LQ + time model has the form γT is spurious.

Actually, of course, as well as their spurious interpretation of their procedures, the analytical procedures themselves by which Travis and Tucker estimated the form for the time factor $f(T) = -\gamma T$ are also quite spurious. The correct statistical procedure to employ to assess the possibility that a given model underfits a given set of data is that of residual analysis - comparing each of the observed responses with that estimated from the model at the same level of the covariates - as we have described extensively in previous sections. If the residuals, say e_i, obtained by comparing the "observed" LD50 values of Table II of the Field et al report with those estimated from the extreme-value model of Travis and Tucker (1986), where $e_i = \ln(-\ln\pi) - [\ln(rN) - \alpha D - \beta D^2/N]$ and using the 12 hr estimates of rN, α, and β from Table VII of that report are plotted against the cognate values of T for the Wara, it is found that the parametric form of the time factor of the LQ + time model could be either $\delta \log T$ or $\delta\sqrt{T}$. We note that a logarithmic form is consistent with our previous conclusion based on our multivariate probit model of the LQ hypothesis for the Tucker and Thames (1983) data (See Section 7.9 and Annex II). But although this conclusion is based upon a correct statistical procedure, little credibility can be put in this estimate either, since it is deployed on weak data. It is simply the best that can be achieved from the sets of data for the studies cited in the Travis and Tucker study.

We have remarked that the first weakness in the procedure used in all three studies (Wara et al, Mah et al, and Travis and Tucker) to combine animal and clinical information is that the respective studies are not parallel and hence the respective models of the clinical and animal data within each cannot be isomorphic: In all three reports the animal studies are of isoeffect while the clinical studies are of dose-response.

Another weakness in the Travis and Tucker report should be remarked in addition to our demonstration that the sample estimates of α/β and γ/β for the Field et al animal isoeffect data are not robust , but are dominated by a single observation at T=53 days. It is that the Travis and Tucker analysis of the Field et al and Wara et al animal data seems to be more than a bit ad hockery: For a linear model of these data regression methods should be (properly) used to estimate the relationship between a single dependent variable and a weighted linear combination of two independent variables with weights to be estimated from the data. Instead, Travis and Tucker have used regression methods to estimate the relationship between what is surely an ad hoc linear combination of two dependent variables (with a priori weights) and a single independent variable. It is of interest to note here that both weaknesses will also to be found in the Fowler 1991 report reviewed below.

It is instructive, therefore, to construct a non-linear regression model of the LQ + time hypothesis on the Field et al isoeffect data:

$$(1/N)D^2 + (\beta_1/\beta_2)D + (\beta_3/\beta_2)T + \{\beta_0 - z(\pi)\}/\beta_2 = 0$$

Inspection of the correlation matrix of the parameter estimate $\hat{\beta}$ suggests that the model is over-parameterized. An obvious re-parameterization gives

$$D(\pi) = [-\alpha_1 + \sqrt{\alpha_1^2 - 4(1/N)(\alpha_2 T + \alpha_0)}\ ]/2(1/N)$$

For these data $\pi = 0.50$ (the LD50). For this model $\bar{R}^2 = 0.978$. The quasi-Newton estimates, $\hat{\alpha}_j(t_j)$, $0 \leq j \leq 2$, of the parameters and precision of estimate, are $\hat{\alpha}_1(t_1) = 0.945(0.836)$ and $\hat{\alpha}_2(t_2) = -1.625(-3.627)$, where $t_j = \hat{\alpha}_j/\sqrt{(Var(\hat{\alpha}_j))}$. The respective 0.95 CL are (-1.575, 3.465) and (-2.625, -0.626). Note that $\hat{\alpha}_1 = \alpha/\beta$ is consistent with the a priori estimate $\hat{\alpha}/\hat{\beta} = 3.0$ Gy. But, it is equally consistent with $\alpha/\beta = 0$. Note also that $\hat{\alpha}_2 = -1.625$ is consistent with $\gamma/\beta = 2.2$ Gy^2/day although the signs are different. α_2 is the ratio, β_3/β_2, of the coefficients in the corresponding dose-response model for which $\beta_3 < 0$ and $\beta_2 > 0$. However, it can be seen from the above equation for $D(\pi)$ that the isoeffect dose will increase with T - as would be expected a priori. (N.B.: If $\hat{\alpha}_1$ is constrained to be 3.0 ($\alpha/\beta = 3.0$) then the sample estimate of the "time factor" obtained from the non-linear regression model above is $\gamma/\beta = -2.214$, which is consistent with the cognate estimate in Travis and Tucker, 1983.) Figure 45g presents a super-position of the isoeffect data ($D_i(0.50)$, N_i, T_i), $1 \leq i \leq n = 13$ of Field et al 1976 and the isoeffect surface corresponding to the above non-linear model of the LQ + time hypothesis. The surface of Fig. 45g, not the curve constructed by Travis and Tucker that is shown in Fig. 45a, is the appropriate geometry for the isoeffect model of the LQ + time hypothesis. (Compare Figs. 10a (and 10b) with Fig. 10c of the correct nonlinear isoeffect model of the LQ hypothesis.)

As we have shown in our analyses of the received (ED and ESD) models of the Mah et al data, there are implicit boundary conditions that must be satisfied by any model that is not simply a mathematical interpolation procedure, i.e., any model that is to describe a realizable process. One such boundary condition is that the model should not predict a realizable level of the response variable at unrealizable levels of the predictor variables, i.e., a positive response should not be predicted by negative levels of dose - or of time. It is evident from Fig. 45g that for the LQ + time isoeffect model levels of $D(\pi) > 0$ correspond to levels of $T < 0$ as well as $T > 0$. The former is nonsensical, suggesting that the LQ + time hypothesis is simply a mathematical interpolation, or smoothing, procedure for these data rather than a plausible radiobiological mechanism.

It is clear that the clinical dose-response curves of Wara et al (See Figs. 1 and 2 of their 1973 report), Mah et al (Fig. 39a), and Travis and Tucker (Fig. 39b) based upon their respective "equivalent single doses" (ED, ED and ESD) represent, per Kuhn, acceptable solutions to puzzles that serve more to distract than to inform the work of their peers. The respective dose-response models represented by these curves do not represent acceptable solutions to the insistent clinical problems that beset the practicing radiotherapist. The genre of equivalent-single-dose-models rarely even represent an acceptable solution to the less-worldly problem of constructing a statistically adequate model of multivariate dose-response data. (The TDF may be the exception).

One reason that these models of Mah et al and Travis and Tucker do not represent the solution to any problem is, of course, that no problem is ever clearly stated: It is difficult (or even impossible) to determine from any of these three reports, either a cogent reason for introducing the isoeffect information on the animal response into any model of the clinical data or what was the use of what could be achieved thereby. For only one example, what is the difference (increase) in the precisions of the estimates of the parameters of the two probit models of the response, z

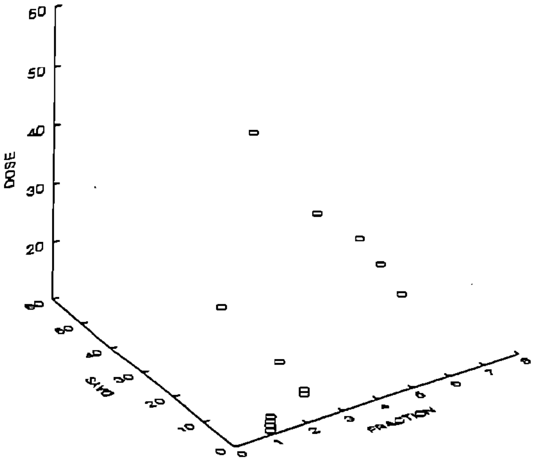


Fig. 45g is a three-dimensional scattergram of the data of the Field et al study. These are $\pi = 0.50$ isoeffect data. A proper geometric model of these data is a surface.

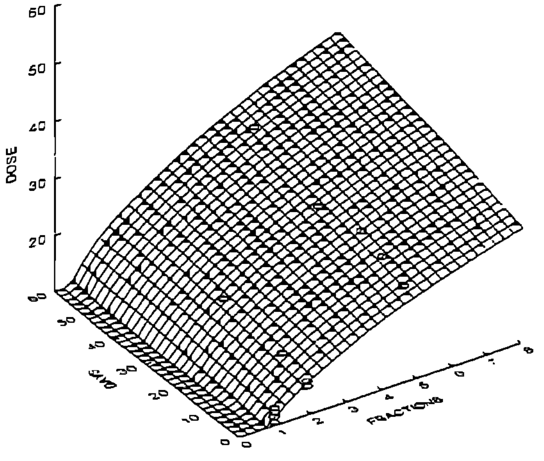


Fig. 45h is a superposition of the $\pi = 0.50$ isoeffect surface for the LQ + time hypothesis estimated on the data of Fig. 45g. The equation of the surface is $D_i(0.50) = [-\alpha_1 + \sqrt{\{\alpha_1^2 - (4/N_i)(\alpha_2 T_i + \alpha_0)\}}]/(2/N_i)$. Note that this is the correct isoeffect model for the LQ + time hypothesis, i.e., a regression of an isoeffect dose, $D(\pi)$ on the predictor variables N and T.

= $\beta_0 + \beta_1 D$ and $z = \beta_0 + \beta_1 ED$? That is, although the incorporation of non-sample information on the parameter must obviously degrade the fit of the model to the clinical data and introduces animal information into the parameter estimates of the clinical model (the resulting model is chimerical - in both the classic and modern senses), there is no discussion of the size of the trade-offs - e.g., bias for variance - that was achieved in the decreased variance of the parameter estimates for the ED model.

Per Kuhn's remark, quoted above, it is obvious that "the conceptual boxes provided by professional education" includes only a simple linear regression "box" (hence the isoeffect models) and a simple probit "box" (hence the equivalent-single-dose dose-response model). Current "professional education" obviously does not include any "boxes" for interval estimates, goodness-of-fit measures and criteria, diagnostics, matrix algebra, etc. It has been one of the principal aims of this report to demonstrate how these deficiencies in professional education affect estimates and inferences and how these deficiencies can be remedied.

In all three of these studies multivariate clinical data were available from which a multivariate dose-response model should have been constructed, say a response surface model, $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where z is the probit or logit (or even the extreme-value) transform of the binary response and the x_1, x_2, and x_3 are, respectively, suitable transforms, say logarithmic, of the dose, D, fractions, N, and (elapsed) time, T, of the treatment regimen. The statistical adequacy of the model could then be assessed by the usual measures of goodness-of-fit (aggregate statistics and diagnostics), precision of parameter estimate ($t_j = \bar{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$ $0 \leq j \leq 3$, and so forth. If the parameter estimates were found to be weak (t_j ~ 1.0, say) then they could, in principal, be strengthened by "mixing", by means of a statistical adequate procedure - say either Bayes or Mixed estimation - in which the information on dose-response that was conveyed by an isomorphic model of multivariate animal data that was obtained in a parallel, that is, a dose-response, study was "mixed" with the information in the clinical sample to obtain posterior estimates of the parameter vector, $\underline{\beta}$. The clinical adequacy of the resulting mixed, or posterior, estimates could then be evaluated. For example, as will be described below, the costs of the increase in the precision of parameter estimates that is achieved by incorporating non-sample (animal) information with sample (clinical) information are explicitly identified and measured. The mixed estimates represent a solution to a problem. The Mah et al and Travis and Tucker reports describe accepted solutions to a puzzle, as will be described further in section 16.6 below. As Kuhn has observed, "Though intrinsic value is no criterion for a puzzle, the assured existence of a solution is."

Table 6a presents the multivariate probit model of the Mah et al data based on the logarithms of the three covariates, D, N, and T for the ungrouped (n=54) data, together with the aggregate (the chi-squared statistic) and case statistics of fit, measures of precision of the parameter estimates and the measures of degradation of predictive performance in future data (the PRESS statistic). The response may be written as $y = \pi(x) + \epsilon$ where the random variable ϵ may assume either of two values: If $y = 1$, then $\epsilon = 1-\pi(x)$ with probability $\pi(x)$; if $y = 0$, then $\epsilon = -\pi(x)$ with probability $1-\pi(x)$. Therefore, ϵ is distributed with mean zero and variance $\pi(x)[1-\pi(x)]$. Thus, the observed response is distributed binomially with conditional mean $\pi(x)$.)

The data obviously do not reject the model on the evidence of the chi-squared statistic. However, there are two observations that are not well-explained by the model and one that dominates the estimates of the parameter vector (See Fig. 40). Recall that Huber (1977) has observed that, "Altogether, 5-10% wrong values in a data set seem to be the rule rather than the exception." Note as well that each of the parameter estimates exceeds its standard error and although $\bar{\beta}_j$ ~ $\sqrt{Var(\bar{\beta}_j)}$, for j = 2,3, we have $\hat{\beta}_1 = 2.76*\sqrt{Var(\hat{\beta}_1)}$, suggesting a "real" - statistically significant - dependence of level of response on level of dose.

Figure 40a suggests that there may be a high degree of multicollinearity present in the (n*k) matrix X of treatment variables (n=54, k=4). This feature of the data is more concisely described by the condition number, $\kappa = \sqrt{\lambda_{(k)}/\lambda_{(1)}}$ where $\lambda_{(k)}$ is the largest and $\lambda_{(1)}$ is the smallest of the eigenvalues of the matrix X^TX (See Belsley, Kuh, and Welsch, 1980). Values of κ around 5 or 10

Table <u>6a</u>. Generalized Linear Model (Probit) of Unaggregated Mah et al (1987) Clinical Data. (n = 54).

$z = -33.235 + 11.944\log D - 3.752\log N - 3.150\log T$
$(-2.439)^{\#}$ (2.761) (-1.038) (-1.001)

Pearson chi-squared. RSS = 50.904 on 50 df.

$P(\chi^2 > RSS|50) = \underline{0.562}$. $\kappa = 197.86^{\#\#}$

PRESS = 58.917.

$^{\#}$ $t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$
$^{\#\#}$ $k = \sqrt{\lambda_{(4)}/\lambda_{(1)}}$

<u>N.B.</u> There are two <u>outlying</u> observations: chi-residuals, χ_i = -2.56 and -2.50. There is one <u>high leverage</u> observation: hat matrix diagonal, h_i = 0.627 (2k/n = 0.148). There is one <u>influential</u> observation: Cook's D_i = 2.457. These are not all distinct; they have index numbers 51, 53, and 54, respectively.

Table <u>6b</u>. Mixed Estimates of $\underline{\beta}$ for Probit Model of Mah et al (1987) Clinical Data. A priori Information from Wara et al Mouse Experiment.

$\underline{r}^T$, R for "Wara model":

$\underline{r}^T = (0,0)$,

$$R = \begin{pmatrix} 0, & 0.377, & 1, & 0 \\ 0, & 0.058, & 0, & 1 \end{pmatrix}, \ \gamma = 1.757,^{\#\#} \quad \theta_p = 0.500.$$

$\psi = 1.0*10^{-8}I_2{}^{\#\#\#}$

$z^{**} = -35.262 + 11.782x_1 - 4.442x_2 - 0.683x_3$. $\hat{\beta}_2{}^{**}/\hat{\beta}_1{}^{**} = -0.377$. $\hat{\beta}_3{}^{**}/\hat{\beta}_1{}^{**} = -0.058$.
$(-2.862)^{\#}$ (2.893) (-2.893) (-2.893)

$^{\#}$ $t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$
$^{\#\#}$ γ is distributed as Pearson chi-squared on 2 df.
$^{\#\#\#}$ I_2 is the 2*2 identity matrix.

<u>N.B.</u> The <u>corrected</u> version of the Mah et al ED model of aggregated data (n=5) required a <u>logarithmic</u> dose metameter. (See Table <u>1c</u>):

$z = -36.789 + 12.284x_1$ where $x_1 = \log ED$
(-2.886) $(2.915)^{\#}$

If the dose-response equation is re-written with $ED = D*N^{-0.377}T^{-0.058}$ then,
$z = -36.789 + 12.284x_1 - 4.631x_2 - 0.712x_3$
where $x_1 = \log D$, $x_2 = \log N$, $x_3 = \log T$. The ratios of these estimates of $\underline{\beta}$ to the cognate mixed estimates, $\hat{\underline{\beta}}$ are, respectively, 1.043, 1.043, 1.043, and 1.043.

indicate that there is no problem with sample multicollinearity; values of κ between 30 and 100 identify moderate to severe degradation of the sample estimates of regression coefficients by the multicollinearity present in the sample. (It will be recalled from sections 6 and 7 that the presence of multicollinearity in sample data will inflate the components of both the sample estimate, $\hat{\underline{\beta}}$, of the parameter vector and the variance-covariance matrix Var($\hat{\underline{\beta}}$).) For the variables of the probit model of Table 6a the condition number, κ = 197.86, suggests that the multicollinearity may severely degrade the parameter estimates of this model. However, it appears that the high level of multicollinearity has only degraded the precision, $\hat{\beta}_j/\sqrt{\mathrm{Var}(\hat{\beta}_j)}$, of these estimates as the respective signs of the $\hat{\beta}_j$, $0 \leq j \leq 3$, appear to be consistent with a priori information. (Note that the condition numbers given in this section are constructed for the cognate linear model $\pi_i = \underline{x}_i^T\underline{\beta} + \epsilon_i$, with the same linear predictor, $\eta_i = \underline{x}_i^T\underline{\beta}$, $1 \leq i \leq n$; that is, they assume a diagonal weight matrix $W = I_n$. We have assumed, based on our previous experience, that the condition numbers for the matrix X^TX will not differ consequentially from the condition numbers of the matrix X^TWX where the weight matrix, W, is that of the appropriate generalized linear model, e.g., probit or logit for the estimate $\hat{\underline{\beta}}$.) Examination of the eigenvectors of the X^TX matrix suggests that it is the high correlation of the variables x_2 and x_3 that may be responsible for the lack of statistical significance of the sample estimates of β_2 and β_3. (See section 7.1). N.B. It should be recalled that in section 7.9, it was shown that the model M1a, which has the same form as the model of Table 6a, provides a much better fit to the experimental data of Tucker and Thames (1983) than did either of the rival models M2 (LQ) or M3 ("LQ + time"). See also Annex II, part 3.

One possible remedy for the effects of multicollinearity is to "shrink" the estimate of $\underline{\beta}$ by the method of Ridge regression (See sections 7.2.2 and 7.11.2 and Figs. 35, 36, and Annex IV). Another remedy is that of Mixed estimation using as a priori information on $\underline{\beta}$ the estimates of the ratios of β_2/β_1(= 0.377) and β_3/β_1(= 0.058) from the Wara et al mouse experiment. (It will be recalled that Ridge regression can be shown to be equivalent to Mixed estimation. See Montgomery and Peck, 1982.) It is evident from Table 6b that the introduction of the a priori information has improved the precision of estimate of β_2 and β_3: the posterior estimates of each are more than twice the respective standard errors. Note, however, that this information carries a specious precision since it requires the assumptions that the within-sample error of estimate of the ratios from the Wara et al data and the between-sample "novel error of uncertain relevance" are both zero. (These are the errors denoted by ϵ_{ij} and δ_{ij}, respectively, in the DuMouchel and Harris study described in section 14 above.) Moreover, the estimates of the ratios of the regression coefficients which are used to convey, or to couple, the a priori, or non-sample, information on $\underline{\beta}$, to the sample information are, as we have remarked in sections 7.3 and 7.4, biased - and the bias may be large if the estimates are not very precise.

Table 6b presents the Mixed, or posterior, estimates, $\hat{\underline{\beta}}^{**}$, of the parameter vector, $\underline{\beta}$, of the probit model of the data in Table 3 of Mah et al (1987) together with a measure of the consistency, γ, of the a priori and sample information and a measure, θ_p = 0.50, of the proportion of a priori information in the posterior estimate, $\hat{\underline{\beta}}^{**}$, of the model. It should be recalled that $\hat{\underline{\beta}}^{**}$ is a matrix-weighted average of a priori and sample information:

$$\hat{\underline{\beta}}^{**} = [X^TWX + R^T\psi^{-1}R]^{-1}\ (X^TWX\hat{\underline{\beta}} + R^T\psi^{-1}\underline{r})$$

$$\mathrm{Var}(\hat{\underline{\beta}}^{**}) = [X^TWX + R^T\psi^{-1}R]^{-1}$$

Note that, on the evidence of the statistic γ, which is distributed as chi-squared on q=2 degrees of freedom, the sample and non-sample, ("Wara model"), information on $\underline{\beta}$ are not inconsistent. However, the clinical data are very weak; that is, they exhibit no strong intrinsic structure, they are "user-friendly" data. For example, they also do not reject the "NSD model" information on $\underline{\beta}$, represented by the constraint.

$$r = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, \ R = \begin{pmatrix} 0, & 0.24, & 1, & 0 \\ \\ 0, & 0.11, & 0, & 1 \end{pmatrix}$$

However, the a priori information on $\underline{\beta}$ provided by the Wara et al isoeffect study can only specify the ratio of two parameters, not the parameters themselves, β_j, $1 \leq j \leq 3$, as would be the

Table 7a. Travis and Tucker Model of the Multifraction LQ Hypothesis on the Mah et al (1987) Clinical Data (n = 54).

N.B. Travis and Tucker did not construct this model. However, it is the model that is entailed in their formulation of the response: "effect level", $E = \pi$.

$$\pi = 5.216*10^{-3}D + 7.993*10^{-3}D^2/N - 9.411*10^{-3}T$$
$$(0.306)\# \quad (1.716) \quad (-0.820)$$

$R^2 = 0.815$. $\bar{R}^2 = 0.714$. $\kappa = 23.19\#\#$. $\alpha/\beta = -.652$ Gy.

$^{\#}\ t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$
$^{\#\#}\ \kappa = \sqrt{\lambda_{(4)}/\lambda_{(1)}}$

N.B. There are two outlying observations: Studentized residuals, $e_9^* = 2.500$, $e_{54}^* = -2.687$. There are two high leverage observations, hat matrix diagonals, $h_9 = 0.440$, $h_{54} = 0.524$. There are two influential observations: Cook's $D_9 = 1.484$, $D_{54} = 2.362$. There are three observations for which $\pi_i > 1.0$: i = 34, 35, and 36. (The subscripts are the row (index) numbers of the observations in the data matrix.)

Table 7b. Generalized Linear Model (Logit) of Multifraction LQ Hypothesis on the Mah et al (1987) Clinical Data (n = 54). $f(T) \propto T$ (Travis and Tucker, 1987.

$$z = -\ 2.2900 + 0.01391D + 0.04253D^2/N - 0.04414T$$
$$(-0.958)\# \quad (0.109) \quad (1.581) \quad (-0.641)$$

Pearson chi-squared, RSS = 50.451 on 50 df

$P(\chi^2 > RSS|50) = 0.456$. McFadden's $\rho^2 = 0.148$.

PRESS = 61.7343. $\kappa = 32.61^{\#\#}$. $\alpha/\beta = 0.327$Gy.

$^{\#}\ t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$
$^{\#\#}\ \kappa = \sqrt{\lambda_{(4)}/\lambda_{(1)}}$

N.B. There are four outlying observations: standardized chi-residuals $|\chi_i^*| \geq 2.0$. There are four high leverage observations: hat matrix diagonal, $h_i \geq 0.148$ (=2k/n). There are two influential observations: Pregibon's $D_i \geq 5.0$. These observations are not all distinct; they have row (index) numbers 9, 51, 53, and 54, respectively.

N.B. The models in Table 7b and Table 7c are non-nested rivals and hence can only be discriminated on the basis of the respective t_j, the diagnostics, and the PRESS (or AIC) statistics, not on the basis of the decrements in deviance (or Pearson chi-squared) statistics.

N.B. The Pregibon D_i statistic used in the logistic analysis of Table 3 differs from the Cook's D_i statistic used in the probit analysis of Table 2 by the factor k^{-1}, where k = p+1, the number of free parameters in the model. Both are measures of the influence of the i^{th} observation on the sample estimate of the parameter vector $\underline{\beta}$ of the model.

Table 7c. Generalized Linear Model (Logit) of the Power-law Hypothesis on the Mah et al (1987) Clinical Data (n = 54).

$$\begin{array}{llllllll} z = & -55.039 & + & 19.767\log D & - & 5.768\log N & - & 5.575\log T \\ & (-2.281)^{\#} & & (2.576) & & (-0.964) & & (-1.054) \end{array}$$

Pearson chi-squared, RSS = 50.91 on 50df. McFadden's $\rho^2 = 0.161$.

$P(\chi^2 > RSS|50) = 0.437$.

PRESS = 58.95. $\kappa = 197.86$.[##]

[#] $t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$

[##] $k = \sqrt{\lambda_{(4)}/\lambda_{(1)}}$

N.B. There are four outlying observations: standardized chi-residuals $|\chi_i^*| \geq 2.0$. There are two high leverage observations: hat matrix diagonal $h_i \geq 0.148$ (=2k/n). There is one influential observation: Pregibon's $D_i \geq 5$. These observations are not all distinct; they have row (index) numbers 9, 51, 53, and 54.

Table 7d. Generalized Linear Model (Logit) of Second Rival Hypothesis on Mah et al (1987) Clinical Data (n = 54).

$$\begin{array}{llllllll} z = & -6.160 & + & 0.111D & + & 1.663D/N & - & 0.042T \\ & (-2.029) & & (1.262)^{\#} & & (1.711) & & (-1.130) \end{array}$$

Pearson chi-squared, RSS = 50.600 on 50 df.

$P(\chi^2 > RSS|50) = 0.419$. McFadden's $\rho^2 = 0.155$

[#] $t_j = \hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$

N.B. There are four outlying observations: standardized chi-residuals $|\chi_i^*| \geq 2.0$. There are four high leverage observations: hat matrix diagonals, $h_i \geq 0.148$ (=2k/n). There is one influential observation: Pregibon's $D_i \geq 5.0$. These observations are not all distinct; they have row (index) numbers 9, 51, 53, and 54.

N.B. The goodness-of-fit of the logit models of Tables 7a, 7b, and 7d was also assessed by the respective Hosmer-Lemeshow statistics that are constructed from the sect of grouped observations in which the grouping variable is the expected response, π_i. All three were concordant on this measure also.

Table 7e. Goodness-of-fit Measures for Generalized Linear Model (Logit) of Mah et al (1987) Clinical Data (n = 54). Comparison of Measures of Concordance. Grouping for Hosmer-Lemeshow Statistic (SYSTAT version 5.0, 1990) is Shown (See Table 7c).

	Statistic	P-Value	DOF
Hosmer-Lemeshow	3.221	0.781	6.000
Pearson	50.914	0.437	50.000
Deviance	57.668	0.213	50.000

Cat.	0.100	0.200	0.300	0.400	0.500
Resp Obs	0.000	0.000	1.000	1.000	5.000
Exp	0.000	0.000	0.527	1.136	5.375
Ref Obs	0.000	0.000	1.000	2.000	7.000
Exp	0.000	0.000	1.473	1.864	6.625
Av. Prob.	0.000	0.000	0.264	0.379	0.448
Cat.	0.600	0.700	0.800	0.900	1.000
Resp Obs	3.000	3.000	2.000	13.000	8.000
Exp	3.223	2.567	3.065	12.772	7.335
Ref Obs	3.000	1.000	2.000	2.000	0.000
Exp	2.777	1.433	0.935	2.228	0.665
Av. Prob.	0.537	0.642	0.766	0.851	0.917

form of the dose-response curve for cell-survival: $m_i = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$, $1 \leq i \leq n$, where m_i is a measure of the number of surviving cells at dose D_i. The logit analysis is presented in Table 7b. Note that, again, these "user-friendly" data do not reject this model either. However, there are some obvious and grave weaknesses in the empirical evidence for LQ + time model that is provided by the Mah et al data. Not only are there several outlying and influential observations, but only one of the parameter estimates exceeds its standard error - and that by not much: $\hat{\beta}_3 = 1.58\sqrt{Var(\hat{\beta})}$. That is to say, none of the sample estimates of β_j are statistically significant - the model does not even describe a statistically significant dependence of the binary response on dose: $\hat{\beta}_1 = 0.109\sqrt{Var(\hat{\beta}_1)}$! But it is known, a priori, that the response must be dose-dependent.

Moreover, it seems most unlikely that this lack of statistical significance of the estimates of the parameter vector of the LQ + time linear predictor is an effect of the multicollinearity in the Mah et al sample, since for this model the condition number is $\kappa = 32.61$ which, is only somewhat in excess of $\kappa = 30$, which suggests that the multicollinearity is only of moderate degree. An alternative - and more likely - explanation is that the form of the linear predictor, η of the LQ + time model completely misspecifies the dose-response relationship in the clinical data of Mah et al.

We have shown earlier that the Travis and Tucker representation, $E = \pi$, misspecifies the form of the distribution of the response. Since it appears likely that the LQ + time model of these data is also misspecified there seems little to be gained by increasing the precision of the parameter estimates by constructing posterior estimates of the parameter vector $\underline{\beta}$ of the model from the a priori information on the dose-response relationship in the mouse that is represented,in the notation of Travis and Tucker, by the constraints $\alpha/\beta = 3$ Gy and $\delta/\beta = 2.2$ Gy^2/day. An old military maxim (from an old militarist, Karl von Clausewitz) is quite apposite here: "Never re-inforce failure"!

The logit model of the Mah et al data with the same linear predictor, $\eta = \underline{x}_i^T\underline{\beta}$, as in the probit model of Table 6a is presented in Table 7c for comparison with the logit model in Table 7b. Note that this model does describe a statistically significant dependence of the response on dose. And for this model, the lack of statistical significance of the sample estimates of β_2 and β_3 can be rather confidently ascribed to the multicollinearity ($\kappa = 197.68$) in the data - as in the case of the cognate probit model of Table 6a. Thus, the lack of significance of the sample estimates of two of the parameters of this model of these data can be accounted for in terms of weaknesses in the data rather than weaknesses in the parameterization of the model. Moreover, there are available, as we have described, two well-precedented salvage maneuvers for such circumstances: Ridge regression and Mixed estimation.

It will be recalled that Robins and Greenland (1986) have recommended that investigators "... choose a model that is consistent with the data and yields parameter estimates consistent with ... prior beliefs. But Table 7 discloses that the LQ + time model of the Mah et al 1987 clinical data does not yield parameter estimates that are consistent with prior beliefs for either the identity link function of Table 7a or the logit link function of Table 7b. For the Mah et al data the LQ + time model would seem to be a "non-starter" on the Robins and Greenland criteria.

The case statistics for the model of Table 7c are presented in Figs. 46a - 46e. The Pearson chi-residual χ_i^* is a standardized residual: $\chi_i^* = \chi_i/\sqrt{1-h_i}$, where χ_i is the Pearson chi-residual and h_i is the hat matrix and $1 \leq i = n = 54$, diagonal for the model. Note that the χ_i^* are the components of the PRESS statistic: PRESS = $\Sigma\chi_i^{*2}$. PRESS, of course, is a measure of the goodness-of-fit, that is the predictive performance, of the model in new data just as the Pearson chi-squared statistic, $\chi^2 = \Sigma\chi_i^2$, is a measure of goodness-of-fit of the model to the original sample. The filled symbols in Figs. 46a and 46b identify the residuals for the responders; the open symbols identify those for the non-responders.

For disaggregated data (in which $r_i = 0$ or $r_i = n_i = 1$, $1 \leq i \leq n$), such as that of the Mah et al study, the Pearson chi-squared distribution is not an adequate approximation to the sampling distribution of the sum of squared residuals, either the Pearson chi-squared residuals, $\Sigma\chi_i^2$, or the deviance residuals, Σd_i^2, unless $\Sigma d_i^2 \sim \Sigma\chi_i^2 \sim (n-k)$, the number of degrees of case if the studies were parallel and thus the models isomorphic. Therefore, the precision of the posterior (mixed) estimate of the coefficient of dose, β_1, for the dose-response model is only slightly increased (t_1 = 2.761 to t_1 = 2.893) by the incorporation of the a priori information on the dose-response relationship in the mouse that is represented in the isoeffect model. The increase in the precision of the estimate of the parameter that was achieved was very modest indeed, even on the (specious) assumption that the precision of the a priori information was, effectively, infinite [$\psi = 10^{-8}I_2$]. In particular, the precision of the posterior estimates of the dose-response model, t_j – 2.9, $0 \leq j \leq 3$, that was achieved by this post-hoc salvage maneuver is much less than that required for a clinically useful model, say t_j = 8-9, by about a factor of 2.5-3.0. But neither the modest increase in precision that was achieved by the use of animal information of fictitiously high precision nor the fact that about half of the information on the clinical dose-response relationship that is described by the ED (or ESD) models resides in mouse data is reported in the Wara et al, Mah et al, and Travis and Tucker papers. There are no disclaimers; only commendations (e.g., "In summary, this prospective study has established a distinct dose-response curve for human pulmonary tissues to fractionated radiotherapy"; "In all respects, this study represents a carefully designed, controlled, executed and documented clinical trial." vide supra).

Travis and Tucker 1987 conclude that their LQ + time model of the Mah et al clinical data demonstrates that the latter are, "... consistent with the concept that tissue response in the lung is determined by the extent of cell killing." One can readily demonstrate, however, that this is not at all the case. The representation of the multifraction LQ model of the Mah et al data that is given by Travis and Tucker in their equation (5) clearly requires that their otherwise unspecified effect level E be identified with the proportion, π, of responders since the constant term, say β_0, that must be included in the linear predictors of probit or logit models of binary responses does not appear in their formulations. (The constant term in these models is necessary to assure that the estimated response, π, is zero when the dose and other covariates are also zero.) An analysis of the linear model that is implicit in the Travis and Tucker formulation of their LQ + time model is presented in Table 7a. The analysis strongly suggests that the Travis and Tucker model misspecifies the form of the predictor, $\eta = \alpha_1 D + \alpha_2 D^2/N + \alpha_3 T$, as well as the form of the response or, more precisely, of the link function appropriate to the response. Table 7a presents a generalized linear model of the LQ + time hypothesis for the Mah et al 1987 data with an identify link function: $g(\pi) = \pi$. The identity link function is inconsistent with a priori information on two counts: 1) The identity link function implies a rectangular distribution of tolerance dose, but it is known a priori that tolerance distributions are unimodal (See Figs. 31a-31c). 2) The identity link implies that for some values of the predictor variables, $\underline{x}_i^T$, there may be estimated responses, $\hat{\pi}_i < 0$ and $\hat{\pi}_i > 1$. Like the rectangular tolerance distribution, such responses are inconsistent with a priori information - in this case the definition of π as a probability. Table 7a disclosed that such responses have, in fact, occurred at three observations in the Mah et al data. Table 7a also discloses that none of the sample estimates of the parameters of the linear predictor exceeds its standard error by more than a factor that is much less than 2. The estimates of two of the three parameters are less than their respective standard errors! This is inconsistent with the prior information that the response is dose-dependent. Moreover, there are also several outlying and influential observations for this model of these data.

The appropriate parameterization of the multifraction LQ model of the clinical binary response data in the Mah et al study is given by a generalized linear model, the multivariate logit, or probit, model

$$z = \beta_0 + \beta_1 D + \beta_2 D^2/N + \beta_3 T,$$

where the respective link functions are $z = \log[\pi/(1-\pi)]$ for a logit model, or $z = \Phi^{-1}(\pi)$ for a probit model; $0 \leq \pi \leq 1$, and $\Phi(.)$ is the Normal distribution function. This model describes tissue response in terms of "cell-killing", since the form of the linear predictor is the generalization (based on the rather dubious assumption that the decrement in cell-survival for a given increment of dose is independent of the position of that increment in a sequence of similar increments as discussed in Annex II, part 3) of the linear predictor for the Poisson model of the LQ hypothesis on the

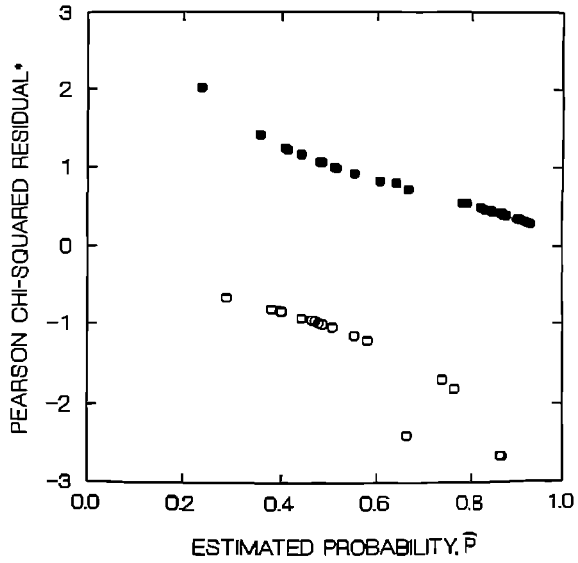


Fig. 46a. The figure presents a plot of the standardized (mean = 0, std. dev. = 1) Pearson chi-residuals, $\chi_i^* = \chi_i/\sqrt{1-h_i}$ vs $\pi_i(=P_i)$ for the model of Table 7c. Here χ_i and h_i are the Pearson chi-residual and hat matrix diagonal, respectively, π_i is the estimated response, and $0 \le \pi_i \le 1$, $1 \le i \le n = 54$.

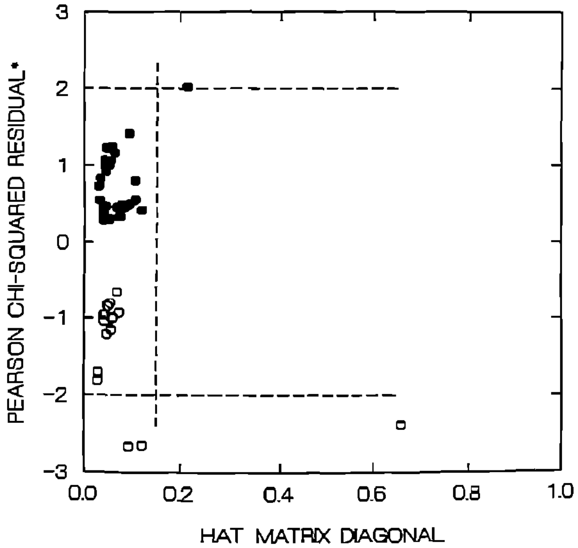


Fig. 46b. The figure presents a plot of the standardized Pearson chi-residual χ_i^* vs the hat matrix diagonal hi for the model of Table 7c. The regions of outlying and high leverage observations lie outside of the three dashed lines. In each figure the fitted symbols identify the responders (radiation pneumonitis) and the open symbols the non-responders.

Comparison of the plots of Figs. 46a and 46b for the model of the non-experimental clinical data of Mah et al (1987) with the respective cognate plots of the experimental animal data of Tucker and Thames 1983 (Figs. 16a, 16b, and 18c) and Thames et al 1982 (Fig. 23b) makes quite vivid the flaws in the experimental designs of the latter two studies. That is, the comparison discloses that the experimental data are no stronger than the non-experimental data, owing to the presence of defects that although unavoidable in observational studies can usually be "designed out" of an experiment.

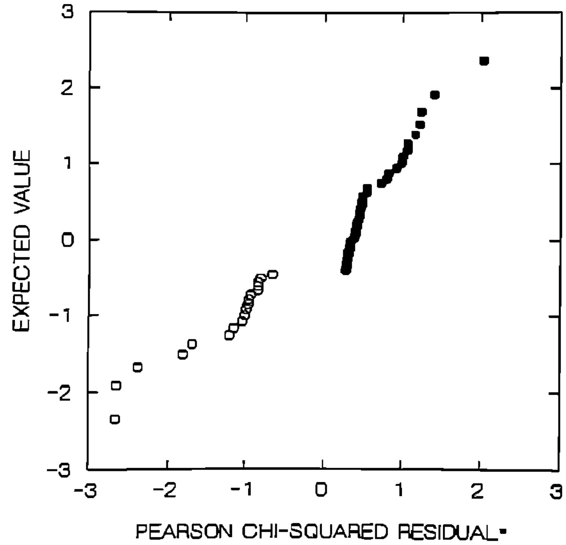


Fig. 46c. The figure presents a Normal probability plot of the distribution of the standardized Pearson chi-squared residual, $\chi_i^* = \chi_i/\sqrt{1-h_i}$ for the model of Table 7c. Symbols are as in Figs. 46a and 46b. The distribution appears to be only approximately Normal, although, as remarked above, it has a mean of zero and unit standard deviation.

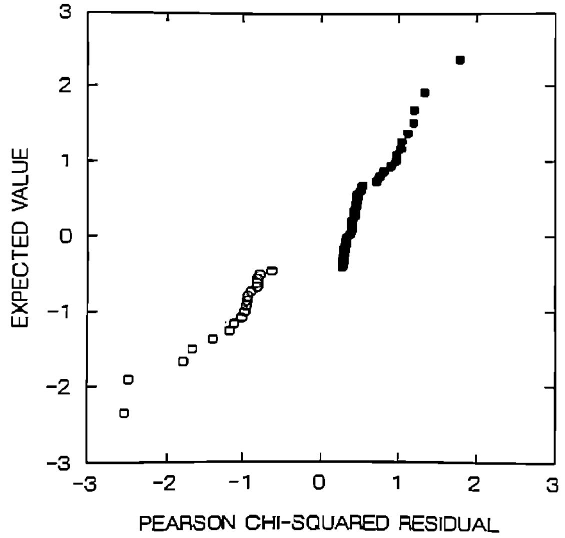


Fig. 46d. The figure presents a Normal probability plot of the distribution of the Pearson chi-squared residual, χ_i for the model of Table 7c. Symbols are as in Figs. 46a and 46b. The distribution appears to be only approximately Normal.

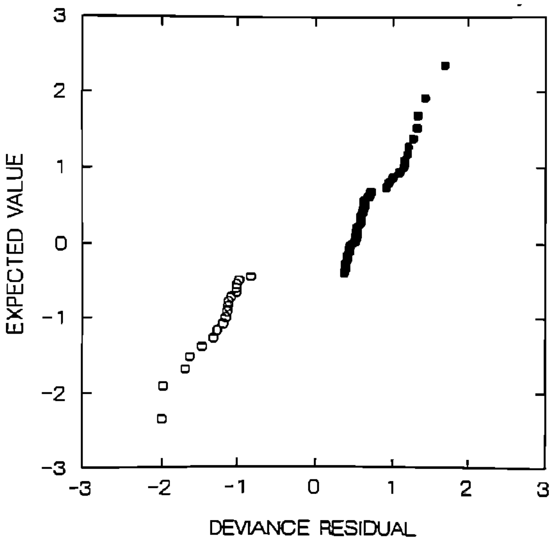


Fig. 46e. The figure presents a Normal probability plot of the distribution of deviance residuals, d_i, for the model of Table 7c. (See Table 2a of section 7.6 of this report. Note that for disaggregated data such as the Mah et al sample for which $r_i = 0$ or $r_i = n_i = 1$, the deviance residuals take the simple forms

$$d_i = -\sqrt{2|\ln(1-\pi_i)} \quad (r_i = 0)$$

and

$$d_i = \sqrt{2|\ln \pi_i|} \quad (r_i = 1)$$

Symbols are as in Figs. 46a and 46b. The distribution appears to be only approximately Normal.

On the criterion of the Kolmogorov-Smirnov (Lilliefors) test the hypothesis of Normality is rejected ($p < 0.001$) for the distributions of all three residuals: χ_i, χ_i^*, and d_i.

Figures 46d and 46e for Binomial data in which $r_i = 0$ or $r_i = n_i = 1$ should be compared with Fig. 18b for Binomial data in which $0 \leq r_i \leq n_i$ and for which the distribution is obviously Normal.

Figures 46d and 46e show that there is little to choose between the Pearson, χ_i, and deviance, d_i, residuals as measures of concordance for disaggregated Binomial data. These plots do suggest that disaggregated data should be binned on the estimated probability, $\hat{\pi}_i$, in order to improve the approximation of the Pearson chi-squared distribution to the sampling distribution of a sum of squared residuals for assessing the goodness-of-fit (As in the Hosmer-Lemeshow procedure).

The standardized Pearson chi-squared residual is the more robust detector of observations that are "not well-explained by the model" since it reduces the effect of the leverage of an observation, measured by h_i, in "masking" it as a true outlier.

freedom, since the expected value of the respective sums is just (n-k).

In general, it is necessary to aggregate - or "bin" - the data such that for most of the "bins" the expected number of responders, $n_j\pi_j$, and non-responders, $n_j(1-\hat{\pi}_j)$ are both on the order 2 or greater. The sum of Pearson chi-squared residuals is now the so-called Hosmer-Lemeshow statistic, C,

$$C = \sum^{g}(r_j - n_j\hat{\pi}_j)^2/n_j\hat{\pi}_j(1-\hat{\pi}_j)$$

where g is the number of "bins", and, for the j^{th} "bin" $1 \leq j \leq g < n$, r_j is the number of responders, n_j is the number at risk, and $\hat{\pi}_j$ is the average estimated probability of response. C is distributed as chi-squared on (g-2) degrees of freedom. It has been shown that the sampling distribution of C is better approximated by the χ^2(g-2) distribution if the "binning" is based on the estimated probabilities, $\hat{\pi}_i$, $1 \leq i \leq n$, where $\hat{\pi}i$ is estimated from the model.

For the model of Table 7c and g=10, the Hosmer-Lemeshow statistic is C = 3.221 on (g-2) = 6 df and $P(\chi^2 > C|8) = 0.781$, i.e., the data do not reject this model. See Table 7e. (Nor is the model rejected on the evidence of $\Sigma\chi_i^2 = 50.91$ on n-k = 50 df. See Table 7c.)

The binning should be done so as to achieve approximately equal numbers at risk, nj, in each range of $\hat{\pi}_i$. Thus, for the model of Table 7c and g=10 we have n_j = 5,5,5,6,6,5,5,4,7,6, where $\Sigma n_j = 54$. It will be recalled that these are about the same as the numbers at risk in most current high dose experiments, e.g., the Tucker and Thames (1983) experiment (Tucker and Thames reported that $n_i = 10$, however, a subsequent check of the actual data disclosed that n_i ¯ 5). Again, a weakness in the non-experimental data emphasizes another of the flaws in current experimental designs: too few subjects at risk at each level of treatment (See Annex II, part 4). The binning described above also gives the following numbers of responders, r_j = 2,3,1,4,3,3,5,2,7,6; $r_j = 36$.

For g=10 there are too many cells (four) for which the expected numbers of responders and non-responders are < 1. Therefore, we use g=5. For this degree of aggregation the expected numbers of responders and non-responders for each cell > 1. The numbers at risk are n_i = 10, 11, 11, 9, 13 and C = 3.546 on g-2 = 3df; $P(\chi^2 > C|3) = 0.315$. (Of course, tests based on the coarser aggregation are less sensitive.)

The non-nested rival logit models

1) $z = \beta_0 + \beta_1\log D + \beta_2\log N + \beta_3\log T$

and

2) $z = \beta_0 + \beta_1\log D + \beta_2 N + \beta_3 T$

of the Mah et al data were compared on the measures and criteria of the respective aggregate and case statistics (including PRESS). Model 1) of Table 7c was the model of choice.

The models of Tables 7b and 7c are also non-nested rivals. Hence they can only be discriminated on the basis of the respective aggregate and case statistics (including PRESS) and the tj statistics. It is apparent that the model of Table 7c is the model of choice for the clinical data of the Mah et al study since $t_j > 1$, $0 \leq j \leq 3$, and the PRESS statistic is the smaller of the two.

Table 7d presents the logit model of another rival hypothesis on the Mah et al data: $z = \beta_0 + \beta_1 D + \beta_2 D/N + \beta_3 T$. The model includes two extensive factors (D and T) and one intensive factor (D/N). If the functional form of the time factor is changed from f(T) = T to f(T) = logT in the models of Tables 7b and 7d, then the estimate of the coefficient of f(T) exceeds its standard error by factors of -1.130 and -0.976, respectively. This suggests that the functional form f(T) = T is incorrect. Note that the curved trend lines of the scattergrams of the Wara et al and Field et al data of Fig. 45b would be made straight by either the log or square root transformation of T. The former would make the animal and clinical information on the form of the time factor more consistent. Comparison of this model with the non-nested model of Table 7b on the criteria of the t_j and PRESS statistics, it would appear that the model of Table 7d is a more adequate parameterization of the dose-response relationship.

Taken together the evidence of Tables 6 and 7 provides a refutation of the multifraction

LQ conjecture (per K. Popper's modus tollens). The Mah et al clinical data may well be, "... consistent with the concept that tissue response in the lung is determined by the extent of cell killing" - although that has not been demonstrated. However, it has been demonstrated that these data are not consistent with the proposition that the LQ + time model provides the best - or even a useful - description of that response.

In Annex II, part 3, it is demonstrated that the generalized linear model (probit) of the experimental radiation toxicity (rat hind-leg paresis) data (Tucker and Thames, 1983) appropriate to the LQ hypothesis does not "fit" these data - on the evidence of the statistically adequate measure, $\Sigma\chi_i^2$ (see also part 7.9.1 and Fig. 16b). However, the sample estimates of the parameter vector are quite precise: $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)} > 6.5$, $0 \leq j \leq 2$. On the other hand, it has been demonstrated above that the generalized linear model (logit) of the non-experimental (clinical) radiation pneumonitis data (Wara et al, 1987) appropriate to the LQ + time hypothesis does fit these data on the evidence of the same statistically adequate measure, $\Sigma\chi_i^2$ (see Table 7b). However, the sample estimates of the parameter vector are quite imprecise: $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)} < 2.0$, $0 \leq j \leq 3$; none of the parameters β_j differs significantly from zero! Taken together these two analyses demonstrate quite vividly that the precision of the parameter estimates cannot be used as a goodness-of-fit criterion, contrary to the practice of Fertil et al (1980): "... the experimental fluctuations and the quality of the fit were expressed by both the variances and the covariance(s) linked to these parameters ... the quality of fitting are represented by a 95% confidence ellipse or ellipsoid ..."

As we remarked in section 1 of this report, the main thrust of our criticisms is not so much to refute certain of the received models of dose-response, as it is to show that certain of the received concepts, methods, and criteria of dose-response modelling are inadmissible. However, one cannot but note that our analyses do indeed tend to refute both the LQ model of the experimental data of van der Kogel (see Annex II, part 3 and Figs. 14-18) and the LQ + time model of the clinical data of Mah et al. This may suggest to some investigators that any decision to deploy the latter model in the clinic should be re-examined in the light of these findings which are based on more statistically adequate concepts, methods, and criteria than have been hitherto exploited in either the original studies or those later studies based on their data. (We remark again on one of the major paradoxes presented by the LQ multifraction model: The response explicitly described by the model is cell-survival, which has a Poisson distribution, yet the model is deployed to explain and exploit the observed responses in tissues for which the distribution is Binomial. Although those investigators who deploy the LQ model hold that such a disjunction provides the correct description and implementation of the hypothesis that "tissue response is determined by cell-killing", in every study that we have examined in which the LQ or LQ + time models were used by the investigators, we found that these models proved to be misspecified - on the evidence of statistically adequate measures and criteria.)

16.6 The Three P's of Radiation Oncology: Paradigms, Puzzles, and Paradoxes

We have noted above that the errors and omissions which we found in the triptych of studies by Wara et al, Mah et al, and Travis and Tucker, that is, the absence of statistically adequate criteria and measures of concordance of model and sample, of consistency of model and non-sample information on the parameter vector, of point and/or interval estimates of the components of the parameter vector and linear functions thereof (quantiles, etc.); the presence of unremarked (and thus, apparently, unnoticed) outlying and influential observations, etc., are all weaknesses that we have also encountered in nearly all of the studies that were reviewed previously; they appear to be generic errors of the peer-group and not just eccentricities of an individual investigator or two (vide infra). We have remarked elsewhere in this report that, taken together, they identify radiobiological modelling as a department of what Fleck (1935/1979) has defined as exoteric, or popular, science: "Popular science in the strict sense is science for the non-experts, that is, for the large circle of adult, generally educated amateurs. ... Characteristic of the popular presentation is the omission both of detail and especially of controversial opinions; this produces an artificial simplification. Here is an artistically attractive, lively, and readable exposition with last, but not least, the apodictic valuation simply to accept or reject a certain point of view. Simplified, lucid, and apodictic science - these are the most important characteristics of exoteric knowledge. In place of the constraint of thought by any proof, which can only be found with great effort, a vivid picture is created through simplification and valuation." As we have observed in an earlier section of this report, it seems that good models are rare <u>because bad models are so easy</u>. As we have also noted elsewhere in this report, one of the functions of statistical methodology, e.g., the measures and criteria of goodness-of-fit, is to provide - or to implement - the constraints of reason on the otherwise often untethered "thoughts" of an investigator. For, as Harris (1975) has remarked, "Statistics is a form of social control over the professional behaviour of researchers." (Perhaps more to the point in the present context is Feinstein's (1990) observation that "The role of scientific methods is to restrain the investigator's advocacy.")

16.6.1 <u>Paradigms</u>.

We have noted above that the Wara et al study has been cited 166 times since publication (to July 1990). This suggests that it has been an extra-ordinarily influential study in this field. (It will be recalled that a recent (1991) study published in <u>Science</u> has reported that less than half of the papers published in major journals are cited even once after publication.) We also remarked that the principal features - including the principal errors - of the Wara et al study were repeated in the subsequent studies of Mah et al and Travis and Tucker. We noted as well that none of these three studies provided any information that would have been useful to either the prediction or the explanation of the observed clinical response. That is, none of the studies provides interval estimates of the lower quantiles, say ED(0.05) or ED(0.10), nor did they provide either point or interval estimates of the parameter vector <u>β</u> of the respective models, nor did they provide statistically adequate estimates of the goodness-of-fit thereof. The fact that our secondary analyses disclosed that the interval estimates of both quantiles and parameters were so wide that neither could have been usefully exploited in their respective roles is for the moment beside the point, which is that none of the three studies attempted to provide useful models conveying useful information. (Of course, as Kuhn has noted, "... an excessive concern with useful problems, regardless of their relation to existing knowledge and technique can so easily inhibit scientific development.")

Taken together the above evidence identifies the Wara et al study, like the more recent Douglas-Fowler study, as one of the <u>paradigms</u>, (Kuhn's concept and locution), that informs and guides certain of the current practices of radiation biology. That is to say, it represents both "the existing scientific tradition" and the "scientific achievement" (attested to by publication in a scientific journal); it is "the shared example" that identifies the acceptable problems which are available to the scientists who choose to work in this field as well as the legitimate criteria for their solution and the admissible methods for achieving it.

Kuhn has remarked on a curious feature of the "paradigm-based research" which sharply distinguishes it from most other intellectual and artistic enterprises: Those who share belief in the paradigm provide the only audience and only judges of its reported performance. For this reason the reports of researchers "... appear as brief articles addressed only to professional colleagues, the men whose knowledge of the shared paradigm can be assumed and who prove to be the only ones able to read the papers addressed to them." Such reports are not "... intelligible to a generally educated audience" - as we have shown repeatedly in this task group report. In this enterprise it is, as we have noted earlier, often <u>more important to the individual investigator that views be held in common than they be correct.</u>

16.6.2 <u>Puzzles</u>.

T. Kuhn, who first re-discovered the work of Ludwig Fleck quoted above and later introduced both the term "paradigm" and the several concepts which it conveys into the contexts of the contemporary philosophy and sociology of science, has described the practice of science under a paradigm as normal science. He has further characterized that enterprise as puzzle-solving: "Under normal conditions the research scientist is not an innovator but a solver of puzzles, and the puzzles upon which he concentrates are just those which he believes can be both stated and solved within the existing scientific tradition." And, "Though intrinsic value is no criterion for a puzzle, the assured existence of a solution is ..."

...

"Closely examined, whether historically or in the contemporary laboratory, that enterprise [normal science] seems an attempt to force nature into the preformed and relatively inflexible box that the paradigm supplies. No part of the aim of normal science is to call forth new sorts of phenomena; indeed those that will not fit the box are often not seen at all. Nor do scientists normally aim to invent new theories, and they are often intolerant of those invented by others. Instead, normal-scientific research is directed to the articulation of those phenomena and theories that the paradigm already supplies."

...

"No theory, model, or hypothesis, is replaced before it has ... ceased adequately to support a puzzle-solving tradition."

...

"... in scientific practice, as seen through the journal literature, the scientist often seems to be struggling with facts, trying to force them into conformity with a theory he does not doubt."[18]

...

"If it is to classify as a puzzle, a problem must be characterized by more than an assured solution. There must also be rules that limit both the nature of acceptable solutions and the steps by which they are to be obtained."

...

"Normal science does not aim at novelties of fact or theory and, when successful, finds none."

...

"Normal science, for example often suppresses fundamental novelties because they are necessarily subversive of its basic commitments."

...

Clearly, a good way to avoid finding novelty is to not look for it - which is perhaps one reason why statistical methodology, concerned as it is with measures and criteria of goodness-of-fit and uncertainty of estimate and inference, might be considered by some investigators to be subversive of the commitments of normal science - and of current radiobiological modelling.

16.6.3 Paradoxes.

"All CT images were obtained using a fourth generation scanner with 1.0 cm slice thickness and a 3.3 sec scan time" (Mah et al, 1987).

...

"Vertical error bars represent binomial standard deviations for each data point. Horizontal error bars represent the standard deviation of the ED values about the mean ED within each group" (Mah et al, 1987).

...

"The goodness-of-fit to the data points was greater than 95% according to the chi-squared. The ED50 was predicted at 1000 ED units with a standard error of ± 40." (Mah et al, 1987).

These three quotations, taken from the same paper appearing in one of the most widely-read radiation oncology journals nicely limn one of the central paradoxes in modern radiation oncology: the concomitant occurrence of highly developed levels of physics, engineering, and technology, which can be currently deployed in patient care, together with the cargo-cult[17] levels of statistical methodology that are regularly brought to bear on the acquisition and analysis of the clinical and laboratory experiences required to better understand, inform, and guide that deployment in ever more fruitful ways.

This paradox is not without its ethical aspects, of course, several of which Altman and others have commented on at length and which we have also discussed briefly in section 10 of this report:

'So what is the relation between statistics and medical ethics?"

...

'Stated simply, it is unethical to carry out bad scientific experiments. Statistical methods are one aspect of this. However praiseworthy a study may be from other points of view, if the statistical aspects are substandard then the research will be unethical. There are two principal reasons for this. Firstly, the most obvious way in which a study may be deemed unethical, whether on statistical or other grounds, is the misuse of patients (or animals) and other resources. As May has said: '... one of the most serious ethical problems in clinical research is that of placing subjects at risk of injury, discomfort, or inconvenience in experiments where there are too few subjects for valid results, too many subjects for the point to be established, or an improperly designed random or double-blind procedure.' Secondly, however, statistics affects the ethics in a much more specific way: it is unethical to publish results that are incorrect or misleading. Errors in the use of statistics may occur at all stages of an investigation, and one error can be sufficient to render the whole exercise useless. A study may have been perfectly conceived and executed, but if it is analyzed incorrectly then the consequences may be as serious as for a study that was fundamentally unsound throughout."

A.G. Altman, 1982

17. Coda.

One of the more distinguished of the several external referees of the TG1 report remarked that, "The message [of the report] is not as badly needed as it was when TG1 was originally set up." This remark could be taken to imply that the message of the report is still "badly needed" - just not "as badly needed", i.e., that there has been a substantial improvement in the statistical quality of the radiobiological literature since the publication date (1987) of the most recent report evaluated by the TG1.

Any change for the better is, of course, greatly desired and is to be applauded wherever - and whenever - it might occur. However, a further review of the recent and current literature suggests that the remarks by our referee greatly exaggerate the statistical quality of the current literature. In support of our finding, we present a summary of our anatomizations of three reports that were published in the peer-reviewed literature since 1987. The studies were published in 1988, 1991, and 1992. Our examination of these reports disclosed the presence of many of the same weaknesses that we identified in the previous (≤ 1987) literature; the pattern of errors identified earlier continues to propagate across the literature. For example, we found: 1) Reported sample estimates and inferences based on an isoeffect model of the LQ hypothesis, constructed and assessed by ad hoc methods from a sample of dose-response data that can be shown, by standard statistical methods, to reject both the LQ +time and Power-law dose-response models (compare Withers et al 1988 with Tucker and Thames 1983). 2) Construction of an isoeffect curve in circumstances where an isoeffect surface is the statistically appropriate geometrical representation required (An ad hoc linear combination of dose and fractions is regressed upon time for a set of clinical data in order to obtain an estimate of the time factor for the LQ + time isoeffect model. Compare Fowler 1991 with Travis and Tucker 1987). 3) Reported sample estimates and inferences based upon a linear regression model of data in which a single observation dominates these estimates and inferences (compare Keane et al 1992 with Thames et al 1982).

17.1 Non-parametric, non-linear, and generalized linear regression analysis of a set of clinical data.

"Let the data speak." A. Ehrenberg (1975)

The first additional (post-1987) study selected for secondary analysis is reported in the 1988 paper by Withers et al describing a complex role for the time factor in modulating the tumor response. (See also Herbert, 1993a and DuMouchel and Herbert, 1993.) The paper has been cited more than 90 times since publication, making it a "classic" of this field. Figure 47, reproduced with

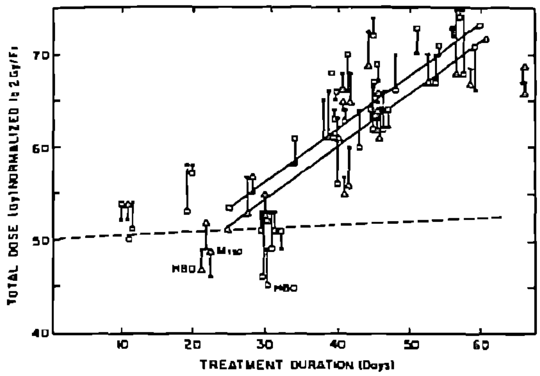


Fig. 47. A scatterplot redrawn from Fig. 1 of the Withers et al 1988 report: "TCD_{50} as a function of overall treatment time for squamous cell carcinoma of head and neck (Table 1). ... Rate of increase in TCD_{50} predicted from a 2 month clonogen doubling rate (---). Estimated increase in TCD_{50} (__) with time for 'T3' (□) and mixed T stages (△) from independent scattergram analyses (Tables 2,3) involving different data sets from those presented in this figure."

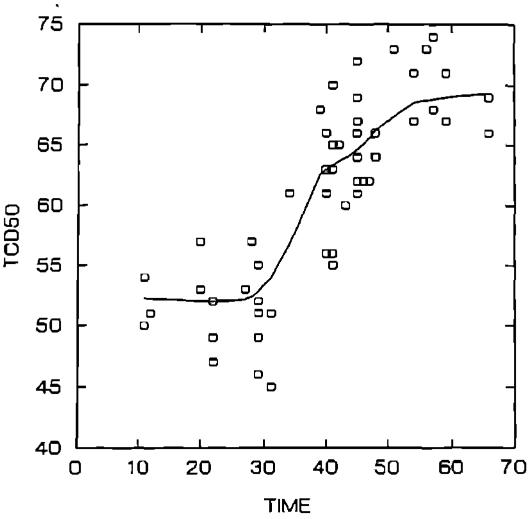


Fig. 48a. Superposition of the scattergram of the data of Fig. 47 and the smoothed curve obtained by Cleveland's 1979 robust locally weighted regression (LOWESS) algorithm. The smoothness of a LOWESS curve is controlled by a parameter f, $0 \leq f \leq 1.0$, which controls the width of the smoothing window, i.e., $f = 0.50$ means that half of the points are included in the running window and contribute to the estimate of the smoothed point $(\hat{y}_k, x_k)$. In Fig. 48a the width of the smoothing window is $f = 0.50$.

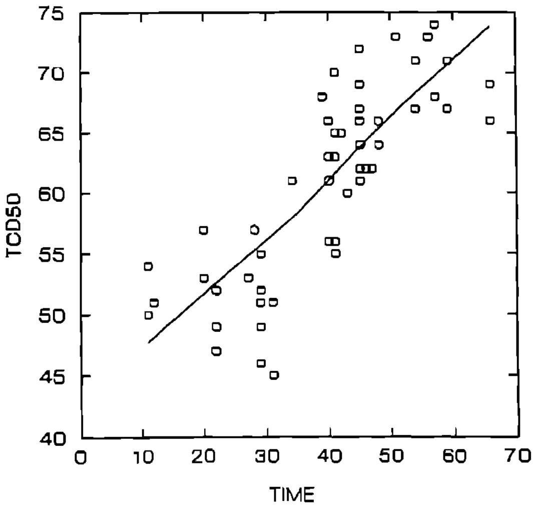


Fig. 48b. Superposition of the scattergram of the data of Fig. 47 and the smoothed curve obtained by Cleveland's Lowess algorithm for $f = 1.00$. This is, effectively, the simple linear regression of TCD50 on time.

permission from their report, has been interpreted as evidence for a "... non-linear, two-component relationship between TCD_{50} and overall treatment time ..." However, Withers et al report that the respective slopes and point of intersection of the two components are <u>not</u> estimated from the sample data described in Fig. 47 but are obtained from other, non-sample, sources; moreover, the concordance of these non-sample estimates with the sample data of Fig. 47 is <u>not</u> formally assessed by any statistically adequate methods, but only "by eye" in the Withers et al 1988 study, a practice we also found in Tucker and Thames, 1983 in which the slope and intercept of each of the seven dose-response curves in their Fig. (our Fig. 9) were estimated "by eye".

Figure 48 describes the curves that we obtained by application of a smoothing algorithm to the data of Fig. 47, or, in the now-preferred usage, by a non-parametric regression. For Fig. 48 the smoothing algorithm is Cleveland's LOWESS. (Cleveland, 1979). However, other smoothing algorithms give qualitatively similar results. Proper selection of the smoothing parameter is critical to the success of all such algorithms. Figure 48a is obtained by choosing the LOWESS smoothing parameter, f, to include 50% of the nearest levels of time for each point estimate of TCD50. Figure 48b is obtained by choosing the smoothing parameter to include 100% of the levels of time in the sample - as in a simple parametric linear regression. "The goal in the choice of f is to pick a value as large as possible to minimize the variability in the smoothed points without distorting the pattern in the data." (Cleveland, 1979). However, the choice of the smoothing parameter obviously also depends strongly on prior beliefs about the nature of the process generating the sample observations - in addition to it being described by a <u>smooth</u> function. Thus, although the use of non-parametric regression does indeed "let the data speak," in the present case, it is necessary for the investigator to select, a priori, the fraction, say f, of the data that speaks to any one point, say (y_i, x_i), $1 \leq i \leq n$, at a time. (The situation is quite analogous to a Cluster Analysis (Anderberg, 1973) in which the investigator must select, a priori, the <u>number</u> of clusters into which the data is to be partitioned by the clustering algorithm - say by k-means.) Figure 48a suggests that a linear spline, or a piecewise linear regression model, may adequately describe the dependence of TCD50 on time. Figure 48b suggests, alternatively, that a simple linear regression model may adequately describe this dependence. If the latter is the case, then the slope of that line is 0.467 with 0.95 CL (0.379, 0.555).

Figure 49 describes the piece-wise linear regression model, $TCD50 = \beta_0 + \beta_1 T + \beta_2 (T-\beta_3)*(T>\beta_3)$, of the Withers et al data of Fig. 47 that was suggested by Fig. 48a. The several parameter estimates, $\hat{\beta}_j$, $0 \leq j \leq 3$, together with the respective 0.95 confidence limits, are: $\hat{\beta}_0$ = 54.06 (43.69, 64.43), $\hat{\beta}_1$ = -0.13 (-0.72, 0.46), $\hat{\beta}_2$ = 0.70 (0.09, 1.30), $\hat{\beta}_3$ = 23.32 (15.24, 31.41). For this model, adjusted $\bar{R}^2 = 0.696$. Note that these estimates from our re-analysis, in which the parameters β_0, β_1, β_2, and β_3 were estimated by an iterative method (as for a non-linear model), are quite consistent with the cognate estimates, obtained by ad hoc methods, that are presented in Withers et al: 54.06 vs 50, -0.13 vs 0.05, 0.70 vs 0.60, and 23.32 vs 25. (<u>N.B.</u>: "Secondary analysis may reanalyze data by ... using competing analytic techniques or sets of assumptions, thus testing the robustness of the original conclusions to alternative approaches. ... If independent reanalyses are done conscientiously and with visibility, the credibility of the original research may be enhanced." T. Hedrick, 1985).

We next "deconstructed" the data represented in Fig. 47 using the information provided in Table 1 of Withers et al 1988. One view of the deconstruction is presented in Fig. 50a, another is in Fig. 50b. In both figures the size (area) of the symbol is proportional to the level of response, π, where $0 \leq \pi \leq 1$. The filled symbol identifies the level $\pi = 0.50$. We fitted the isoeffect model of the Power-law hypothesis to the deconstructed data: $\log D(\pi) = \alpha_0 + \alpha_1 \log N + \alpha_2 \log T$. This gave the parameter estimates $\hat{\alpha}_j$ and the ratios $t_j = \hat{\alpha}_j / \sqrt{(Var(\hat{\alpha}_j))}$, $\alpha_0 = 1.233$, $\alpha_1 = 0.234$, $\alpha_2 = 0.137$, $t_0 = 45.0$, $t_1 = 12.0$, $t_2 = 7.321$, with adjusted $\bar{R}^2 = 0.878$. The model may also be written as, $D(\pi) = 17.11 N^{0.23} T^{0.14}$. The case statistics identified 3 high leverage observations and 1 outlier. However, deletion of these observations did not substantially alter the sample estimates of either the α_j, $0 \leq j \leq 2$, or $\bar{R}^2$. We note that the above estimates $\hat{\alpha}_j$, j = 1,2 of the isoeffect model <u>do not reject</u> the familiar NSD model of normal tissue tolerance.

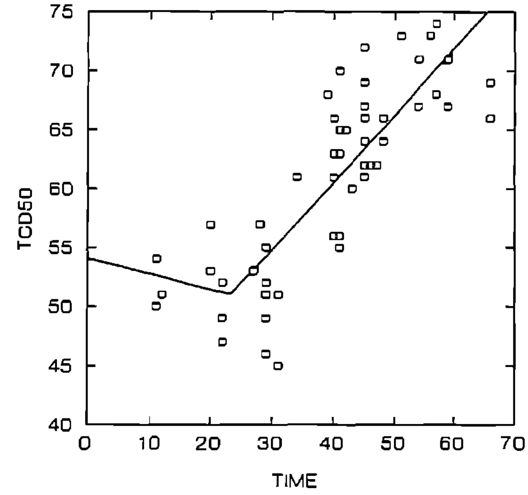


Figure 49. Superposition of the scattergram of the data of Fig. 47 and the piecewise linear regression curve in which the point of intersection ("knot") as well as the respective slopes of the two components are estimated by non-linear (Gauss-Newton) regression methods. The equation of the curve is TCD50 = $\beta_0 + \beta_1 T + \beta_2(T - \beta_3)*(T > \beta_3)$.

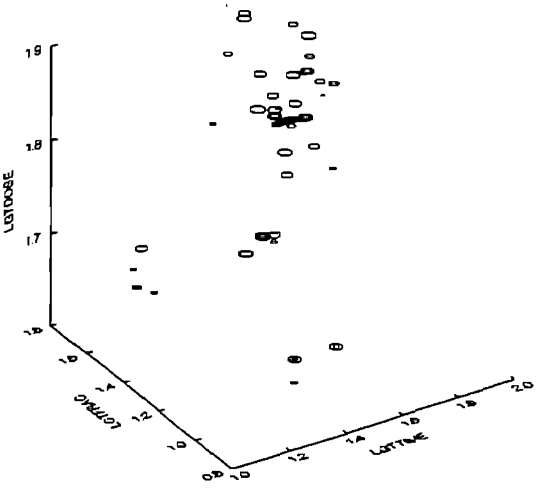


Fig. 50a. A 3-dimensional scattergram of the data in Table 1 of the Withers et al 1988 report. The size (area) of each symbol identifies the proportion, x_i, $0 \le x_i \le 1$, $1 \le i \le n = 59$ of responders at each treatment regimen. The filled symbol in the group of four at lower left identifies the regimen at which $x_i = 0.50$.

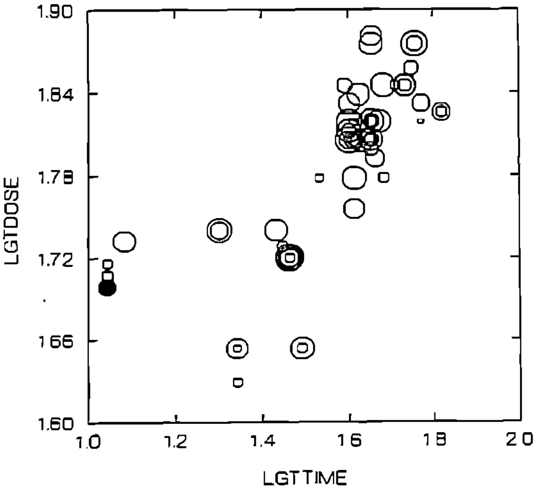


Fig. 50b. Projection of the 3-dimensional scattergram of Fig. 50a onto the logdose-logtime plane. The symbols are as in Fig. 50a.

We next fitted logit dose-response models of both rival hypotheses - Power-law and LQ + time - as described in section 7.5.1.2 above, to the raw data of Table 1 of the Withers et al data. On the evidence of a Pearson chi-squared statistic the data rejected both models with nearly the same levels of α. For the Power-law model the term $\beta_2$logN was not significant. For the LQ + time model the term β_2Dd ($=\beta_2D^2/N$) was not significant. Note that in neither model was the effect of fractionation (N) significant. It is also of interest to remark that evidence for the old "cube-root" law, $D = kT^{0.33}$, can be recovered from the isoeffect model that is embedded in logit dose-response model of the Power-law hypothesis on these data, that is, $\hat{\beta}_3/\hat{\beta}_1 = -$ 0.320, where $\hat{\beta}_3$ and $\hat{\beta}_1$ are, respectively, the estimates of the coefficients of logT and logD in the logit model of the Power-law hypothesis.

Figure 50 presents scattergrams of a deconstruction of the data in which the area of each symbol is proportional to the level of (binomial) response, $0 \leq \pi_i \leq 1.0$, $1 \leq i \leq n = 59$; the filled symbol identifies the $\pi_i = 0.50$ level of response. It is evident that at many of the treatment regimens there are several quite different levels of response evoked by the same level of treatment. Our earlier analyses, showing that the n=59 observed levels of dose, time, and fractions could be adequately described ($\bar{R}^2 = 0.878$) by the isoeffect model $D(\pi) = 17.11N^{0.23}T^{0.14}$ Gy, suggests that the level of treatment for each group was selected on the NSD model of the "tolerance dose" of normal tissue within every group and then the level of tumor response achieved thereby was determined, in part, by other factors such as stage, anatomical site, etc., i.e., the patients in each group were treated to "tolerance". Thus, logit models of both rival hypotheses, LQ + time, and Power-law, were augmented with categorical variables describing the stage of disease, S, and anatomical subsite, A, respectively, based on the data reported in Table 1 of Withers et al 1988. Although the terms β_4S and β_5A were statistically significant for both the Power-law and LQ + time dose-response models, both of the thus augmented models were again rejected by the data. And again, the terms $\beta_2$logN and β_2Dd ($=\beta_2D^2/N$) were both non-significant in the augmented rival models. Note, however, that the addition of terms in stage, S, and subsite, A, did not improve the "fit" of the isoeffect model $D = 17.11N^{0.23}T^{0.14}$ (indeed, these additions slightly degraded the fit, $\bar{R}^2 = 0.874$ vs 0.878); the coefficients of the terms S and A did not differ significantly from zero for the isoeffect model.

(N.B.: It is now a commonplace that hospitals and other health organizations may differ greatly in the "outcome" (the patient's response to treatment) achieved in similar cohorts of patients undergoing similar treatment. This observation has provoked a large amount of "outcomes research" recently. (See Roper et al, 1988; Ellwood, 1988; and Blumberg, 1986). The "outcome" achieved for a given treatment in a given cohort depends on such factors as case mix, case severity, and include measures of, or proxies for, the patient's physiological reserve, etc. (The "risk-adjustment" of the "outcome" for a given health-care provider is based upon such factors.) For the Withers et al, 1988 data the measures of stage and site provide only part of such information. Hence, neither of the rival models provide an adequate "fit" to these data.)

Since the Withers et al 1988 data reject the dose-response models of both of the rival hypotheses, the information on treatment variables and covariates presented in Table 1 of that report does not account for all of the variation in the levels of tumor response presented therein. When insufficient information, e.g., covariates, is available to fully account for the observed variation in response, the model is mis-specified. Since biases, in the parameter estimates and functions thereof, arise when it is impossible to represent a set of data by a model of the type fitted (See "aliasing" in Montgomery and Peck 1982), one should therefore be cautious in deploying the conclusions presented in that paper that are based on the assumption of the validity of the LQ + time model. In particular, the principal conclusion of the Withers et al study, namely, that the time factor is a linear spline with knot at $t \simeq 25$ days, must be taken as unsupported by their data - despite our findings as described in Fig. 48a and 49. (N.B.: The Withers et al study is examined at greater length in the paper by DuMouchel and Herbert, 1993, "Combining Information from Multiple Dose/Time Fractionation Studies. (Bayesian Hierarchical Meta-Analysis Revisited)".) Bentzen and Thames 1991 have also examined the Withers et al 1988 report.

17.2 <u>Isoeffect surfaces. Estimator shrinkage.</u>

"For estimating the coefficients in linear regression, the practice of shrinking the least-squares estimates of the coefficients toward a point, or more generally toward a subspace, has received increasing attention."

J. Rolph, 1976

We next examined the report of a recent study by Fowler (Fowler 1991). This study unfortunately deploys a subset of the flawed concepts, methods, and criteria used in Travis and Tucker 1987 and Withers et al 1988 in an assessment of the putative role of a time factor in modulating the response of normal tissues to fractionated radiation exposures in treatment for head and neck cancer. Figure 51 is reproduced with permission from Fowler 1991. The group of 18 studies represented therein was partitioned by Fowler into one subgroup of 7 ("moderate dose") and one subgroup of 11 ("high dose") and his primary analyses were based on this partition. However, on close examination we found this partition to be a "distinction without a difference" and our secondary analyses were therefore based on the full group of 18 studies. We must also remark on a weakness in these study data that is exacerbated by the partition into the subgroups described in Fowler 1991. This can be shown as follows. Figure 52a describes the "moderate dose" subgroup of Fig. 51. It is obvious from simple inspection that the estimate of the slope (66.1 ± 12.5) in Fig. 51 will be <u>dominated</u> by the single observation at T = 11 days. This is confirmed by the simple analysis described in Fig. 52b in which the size of each symbol is proportional to its influence (determined by deletion) on the product-moment correlation coefficient and hence on the linearity of the relationship (Compare with Fig. 45a above for the Field et al 1976 data). The regression diagnostics for this observation on the Fowler isoeffect model, $[D + Dd/(\alpha/\beta)] = \alpha_0 + \alpha_1 T$,of these data are consistent with graphical analysis shown in Fig. 52b: Cook's $D_1 = 2.62$, $h_1 = 0.66$, $e_1^* = -2.13$ where e_1^* is the Studentized residual. Deletion of the observation at T = 11 days gives a slope estimate of 41.6 ± 15.2 - a change of nearly 40%. It would appear that the published slope estimates for "low dose" subgroups are <u>not</u> robust - an additional reason to <u>pool</u> the "high dose" and "moderate dose" subgroups. (This is an instance of using Data augmentation to overcome a weakness - influential observations - in the sample. See section 7.2.3 above.)

The Fowler 1991 model of the group of 18 observations of clinical isoeffect in Fig. 51 is based upon a kind of transmogrification of the isoeffect model of the LQ + time hypothesis that is quite similar to that found in Travis and Tucker 1987: $D + Dd/(\alpha/\beta) = \alpha_0 + \alpha_1 T$, with $d = D/N$ and $\alpha/\beta = 8$ Gy selected a priori. (In the Travis and Tucker model of the Wara et al, 1973 and Field et al, 1976 mouse isoeffect data we have $D(\alpha/\beta) + Dd = \alpha_0 + \alpha_1 T$ with $\alpha/\beta = 3$ Gy selected a priori.) Note that in the Fowler 1991 and Travis and Tucker 1987 procedures an arbitrary (a priori) linear combination of two dependent variables is regressed upon a single independent variable. However, the isoeffect model appropriate to the LQ + time hypothesis for binary response data is a non-linear model. That is, a single dependent variable, $D(\pi)$, is regressed upon a nonlinear combination of two independent variables, i.e., a proper regression model. Note also that the geometrical representations of the Fowler (and Travis and Tucker) isoeffect models are <u>isoeffect lines</u> whereas that for the (correct) nonlinear isoeffect model, $D(\pi) = f(N, T)$, is an <u>isoeffect surface</u>. We shall return to this issue below.

For the Fowler 1991 isoeffect model of the LQ + time hypothesis on the pooled data (n=18), we have $\alpha_0 = 56.768$, $\alpha_1 = 0.652$, $t_0 = 17.43$, and $t_1 = 7.26$, where, as above, $t_j = \hat{\alpha}_j/\sqrt{(Var(\hat{\alpha}_j))}$, j = 1,2, and adjusted $R^2 = 0.753$. For the isoeffect model of the Power-law hypothesis, $\log D(\pi) = \alpha_0 + \alpha_1 \log N + \alpha_2 \log T$, we have $\alpha_0 = 1.229$, $\alpha_1 = 0.208$, $\alpha_2 = 0.168$, $t_0 = 26.65$, $t_1 = 6.49$, $t_2 = 6.19$, and (adjusted) $\bar{R}^2 = 0.903$. We note that the above estimates, $\bar{\alpha}_j$, $0 \leq j \leq 2$, or $D(\pi) = 16.96N^{0.208}T^{0.168}$, are consistent with the parameters of the familiar NSD model, $D(\pi) = 18.00N^{0.24}T^{0.11}$. That is, for the hypothesis, H_0: $\alpha_0 = 1.255$, $\alpha_1 = 0.24$, $\alpha_2 = 0.11$, we have P = 0.059 for the Fowler 1991 data.

It is of interest to remark that the Fowler 1991 data provide an example of the effect of using Data augmentation to reduce the degree of multicollinearity in a sample. The NSD model is

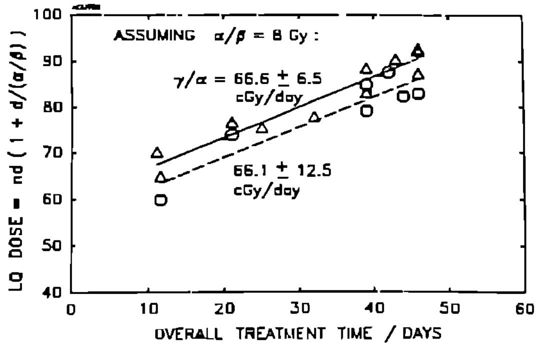


Fig. 51. A scatterplot redrawn from Fig. 1 of the Fowler 1991 report: "The LQ Dose, i.e., the Biologically Effective Dose without a time factor: nd x (1+d/(α/β)) for each schedule in Table 1 as a function of the overall time. _ standard errors of the mean slope are given. Circles 'moderate' dose levels. Triangles 'high' dose levels. The value α/β = 8 Gy was assumed." (Reprinted with permission from Fowler (1991).

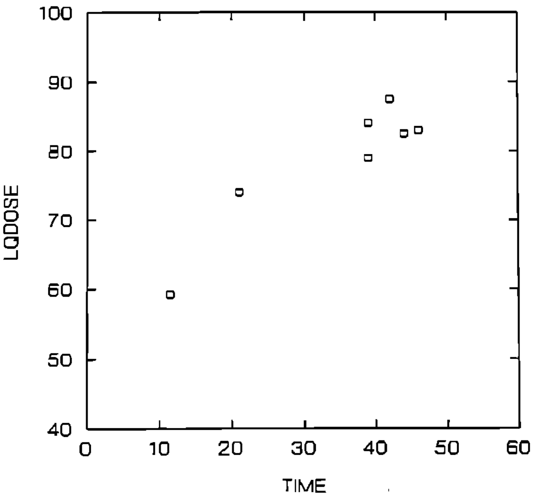


Fig. 52a. Scattergram of the "moderate" dose levels of Fig. 51. The figure suggests that the estimate of the slope will be dominated by the extreme observation at T = 11.5 days.

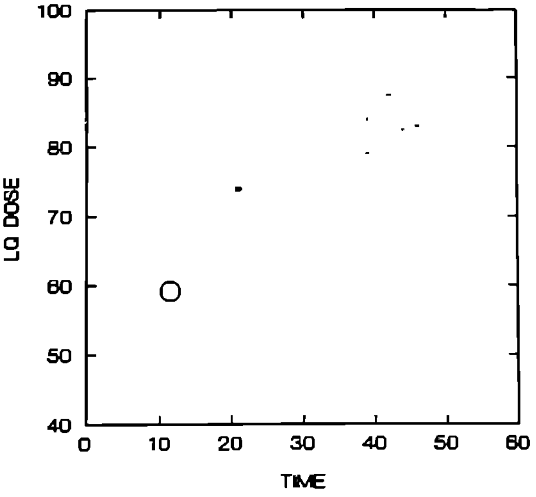


Fig. 52b. Scattergram of the row-deleted diagnostic, R - R(i), where R is the Pearson correlation coefficient for the data of Fig. 52a. The size of each symbol is proportional to the decrement R -R(i) for the i^{th} observation. The symbols are as in Fig. 45f. Obviously, the linearity of the relation, and estimate of the slope, is dominated by the observation at T = 11.5 days.

more consistent with the subset $T > 12$ days of the Fowler data than with the full data. For this subsample of n=15 the estimated isoeffect model of the Power-law hypothesis is $D = 19.79N^{0.26}T^{0.08}$. However, the degree of multicollinearity is high; the condition number is $\kappa = 54.12$, and examination of the eigenvectors of the correlation matrix, X^TX, discloses that the variables logN and logT are highly correlated. This explains the observed inflation of the variance of the estimate of α_2, the coefficient of logT in the Power-law model of the subsample: $t_2 = 0.833$; that is, the parameter estimate, $\hat{\alpha}_2$, is less than its standard error, $\sqrt{Var(\hat{\alpha}_2)}$. However, for the augmented sample (n=18) the condition number $\kappa = 20.74$ and all of the parameter estimates greatly exceed their respective standard errors, evidence that the degree of multicollinearity and its characteristic effects on the parameter vector have been markedly reduced by the addition of the observations with $T < 12$ days.

Examination of the respective sets of case statistics for the two rival isoeffect models (Power-law and LQ + time) discloses the presence of a single common outlying observation. Deletion of this observation gives the following estimates for the isoeffect model of the Power-law hypothesis on the reduced (n=17) data sets, $\alpha_0 = 1.262$, $\alpha_1 = 0.179$, $\alpha_2 = = 0.179$, $t_0 = 32.05$, $t_1 = 6.33$, $t_2 = 7.95$, and adjusted $\bar{R}^2 = 0.930$. (The model may be written as $D(\pi) = 18.29(NT)^{0.18}$.) The above estimates $\hat{\alpha}_j$, $j = 1,2$ are consistent with H_0: $\alpha_1 = 0.158$, $\alpha_2 = 0.158$. Nor do they reject the hypothesis H_0: $\alpha_1 = 0.180$, $\alpha_2 = 0.140$.

These last two sets of estimates are the non-least squares estimators previously obtained by Herbert 1981, 1989d using principal components (PC) and ridge regression (RR) methods, respectively, to obtain parameter estimates for the isoeffect models of the Power-law hypothesis from the data of two previously published reports on clinical studies on head and neck cancer (see Fig. 4b) and on cervical cancer (see Fig. 4a) in which there was a high degree of multicollinearity in the isoeffect data of each of the original studies. (N.B.: It must be remarked, however, that the Ridge regression estimators for the isoeffect model of the Power-law hypothesis on the head and neck cancer data cited above were not consistent with the Fowler 1991 data, a fact that we attribute to sampling error.) The multicollinearity in the last two sets of data inflated both the least squares (LS) estimators of the parameters and their covariance matrix for isoeffect models of the Power-law hypothesis for these earlier studies. It also caused one of the LS parameter estimates in each model to have the "wrong" - inconsistent with prior information - sign as well. For estimating the parameters of regression models of data in which either the degree of multicollinearity is high (as in the present case) or the ratio p/n is high (p is the number of predictor variables and n is the sample size) the RR and/or PC methods are preferred (on both mean-square error and the Robins and Greenland 1986 criteria quoted above) to the LS estimators. See also section 7.2.2 above.

Non-least squares estimators of model parameters such as RR and PC belong to the broad class of shrunken estimators. Figure 35d presents the LS, PC, and RR estimates for the head and neck cancer data referred to above. (N.B.: The latter isoeffect data were disaggregated whereas, the Fowler data are aggregated.) The shrunken estimators offer improvements - in some sense, usually mean-squared error, and in some cases -on the usual least squares (LS) and maximum likelihood (ML) estimators for models of data in which multicollinearity is present. (N.B.: If the true, but unknown, value of a parameter is θ and an estimator of θ is $\hat{\theta}^*$ then the mean-squared error of $\hat{\theta}^*$ is $E(\theta - \hat{\theta}^*)^2$, where E denotes the expected value.) It should be noted that the predictive performance of a given model in extrapolation (i.e., in "new data") should be better for the shrunken estimator (i.e., RR or PC) of the parameter vector than for the LS estimator, since the effects on the sample estimates of the parameter vector of the high degree of multicollinearity - an idiosyncracy - of the construction sample data - are reduced by the post-hoc shrinkage procedures.

Since the degree of multicollinearity is low in the pooled data of the Fowler 1991 study (owing to the presence of the three observations at $T < 12$ days), the LS estimators are optimal for the model parameters. Thus, the parameter estimates of the isoeffect models of the Power-law hypothesis constructed on data on both tumor and normal tissue radiation responses at three different anatomical sites in studies by three different groups at three different times and by three different methods (LS, RR, and PC) are quite consistent for the two variables fractions, and time. A common Power-law model is suggested, say $D = D_0N^{0.16}T^{0.16}$.

The close agreement of the estimates of the parameters of the isoeffect model of the Power-law hypothesis obtained from three quite disparate studies may seem quite remarkable. However, to some, it suggests that many of the so-called isoeffect equations to be found in the literature are in fact "management equations". That is, they describe the philosophy of management of malignant disease by the use of ionizing radiation that prevailed during the time that the data were accumulated. Thus, it more accurately summarizes one aspect of the "standard of practice in the community" than it does the radiation response of normal tissues within the target volume. It describes the intensity and duration of treatment that will evoke the binary response, "Tolerance!" in the attending oncologist. More precisely, the so-called "tolerance dose" refers to an unspecified upper quantile, say 0.95, of the distribution of "tolerance" in the radiation oncologist. It is more of a measure of a psychological response - risk aversion - rather than of a physiological response - tissue reaction. The prevailing philosophy of management has hitherto been to "treat to tolerance" in every case and accept whatever level of tumor control results. This interpretation of our findings on the common Power-law model, $D = D_0N^{0.16}T^{0.16}$ finds support in the views of an eminent radiation oncologist: "The usual approach to radiotherapy practice is to decide on some maximum acceptable probability of injury appropriate to the condition treated and to irradiate to the level of effect with which that probability is associated, knowing that at that level of effect there is also some probability of benefit. This maximum acceptable probability of injury appropriate to the condition treated cannot be related formally and quantitatively to the level of effect; it is related by consensus - the consensus of patient (what is acceptable to the patient), of physician (his judgment and the influence of his peers), and of society (medical legal standards of acceptability)" Andrews, 1982. Thus, our finding that the parameters of the respective management equations for several different disease sites take quite similar values should come as no great surprise since the level of risk that will be tolerated by the oncologist is similar for every site. (Note that Fig. 50 provides a vivid description of this practice of irradiating "to tolerance" and then accepting whatever "probability of benefit" - say, probability of tumor response - results. In the Withers et al 1988 data described in that figure the "probability of benefit" varied between 0.18 and 0.86, and depended only in part on the stage and anatomical sub-site of the disease.)

The foregoing secondary analyses suggest that the Power-law hypothesis provides the isoeffect model of choice for the Fowler 1991 "tolerance" data. These data (n=18) are shown in a 3-dimensional scattergram in Fig. 53. (The filled symbol identified that observation which is an "outlier" for both the Fowler LQ + time isoeffect model and the Power-law isoeffect model.)

Although the Power-law hypothesis is the hypothesis of choice for these data - on statistically adequate criteria - it is nevertheless of interest to show that one can make a stronger argument for the LQ + time hypothesis on the basis of the non-linear isoeffect model than that which was presented in Fowler 1991 on the basis of the received model, $D + Dd/(\alpha/\beta) = \alpha_0 + \alpha_1T$, in which the value of α/β was assigned a priori. The logit (say) dose-response model of the LQ + time hypothesis on a sample of binary data of size n can be written as

$$z_i = \beta_0 + \beta_1D_i + \beta_2(D^2/N)_i + \beta_3T_i.\ 1 \leq i \leq n$$

where $z_i = \ln[\pi_i/(1-\pi_i)]$, the logit transform of the proportion, π_i, of responders at (D_i, N_i, T_i). For a constant level of response - an isoeffect - we have $z_i = z^*$, a constant (corresponding to, say, $\pi^* = 0.05$ for "tolerance") and the equation

$$\beta_2(D^2/N)_i + \beta_1D_i + \beta_3T_i + \beta_0 - z^* = 0$$

defines the set of values of dose, fractions and time that yields the commonly accepted level of response, π^*. The isoeffect model is

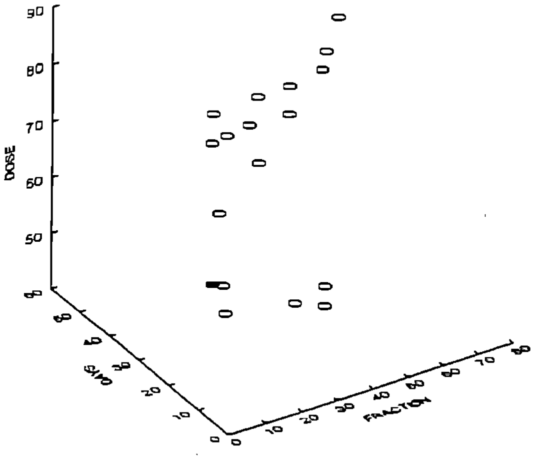


Fig. 53. 3-dimensional scatterplot of the n=18 (isoeffect) data in Table 1 of Fowler 1991. The filled symbol identifies the observation that is an outlier on both the LQ + time and Power-law models of these data. This observation was omitted from the sample from which the isoeffect surfaces of Fig. 54a and Fig. 54b were estimated.

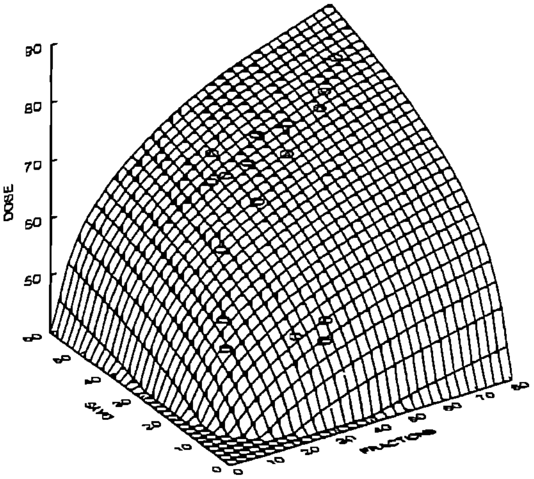


Fig. 54a. Superposition of the reduced (outlier deleted) isoeffect data of Table 1 of the Fowler 1991 report and the Power-law isoeffect surface, $\hat{D}_i(x)$, where

$$D_i(x) = \alpha_0 N_i^{\alpha_1} T_i^{\alpha_2}, \ 1 \leq i \leq n = 17$$

and

$\alpha_0 = 18.29$ Gy, $\alpha_1 = 0.179$, $\alpha_2 = 0.179$.

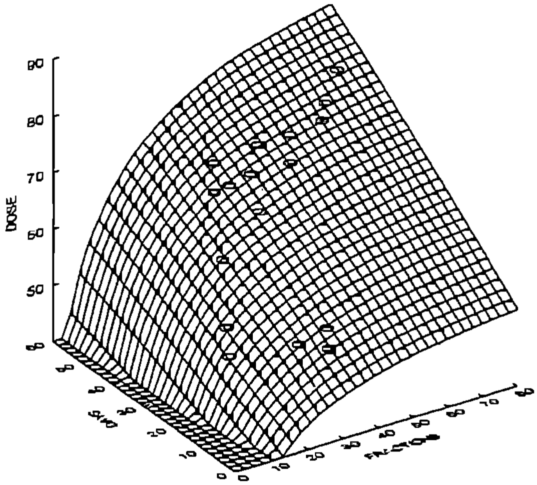


Fig. 54b. Superposition of the reduced (outlier deleted) isoeffect data of Table 1 of the Fowler 1991 report and the LQ + time isoeffect surface, $\hat{D}_i(x)$, where

$D_i(\pi) = [-1 + \sqrt{\{1 - (4\gamma_1/N_i)(\gamma_2 T_i + \gamma_0)\}}]/$
$(2\gamma_1/N_i)$, $i \leq i \leq n = 17$

and

$\gamma_0 = -58.631$, $\gamma_1 = 0.147$, $\gamma_2 = -0.680$

$$D(\pi^*) = \frac{-\beta_1 + \sqrt{\beta_1^2 - 4(\beta_2/N)(\beta_3 T + \beta_0 - z)}}{2\beta_2/N}$$

However, the parameter covariance and correlation matrices showed that this model is over-parameterized (Bates and Watts, 1988; Ratkowsky, 1990). Hence, we used the re-parameterized model,

$$(\gamma_1/N_i)D^2_i + D_i + [\gamma_2 T_i + \gamma_0)] = 0,\ 1 \leq i \leq n,$$

where $\gamma_1 = \beta/\alpha$, $\gamma_2 = \gamma/\alpha$.

An iterative (Quasi-Newton) nonlinear regression procedure yields the point estimates (and 0.95 CL): $\hat{\gamma}_0 = -61.080(-75.244, -46.916)$, $\hat{\gamma}_1 = 0.180(0.060, 0.300)$, $\hat{\gamma}_2 = -0.705(-0.913, -0.496)$; $\bar{R}^2 = 0.906$. A first-order (<u>not</u> delta) approximation to α/β is $1/0.18 = 5.556$, with 0.95 CL (16.667, 3.330). If we fix $\gamma_1 = 1/(\alpha/\beta) = 1/8$, then the <u>constrained</u> estimates are $\hat{\gamma}_2 = -0.688(-0.869, -0.507)$ and $\hat{\gamma}_0 = -55.556(-62.273, -48.837)$. It will be noted that these estimates of α/β (= 5.556 Gy) and γ/α (= -0.795 Gy/day) are different from - but consistent with - the cognate estimates published in Fowler 1991: $\alpha/\beta = 8$ (a priori estimate), $\gamma/\alpha = 66.6$ (sample estimate).

However, all of these parameter estimates are exceedingly fragile. If we delete the three observations for which $T < 12$ days then $\gamma_0 = -90.762(-142.021, -39.504)$, $\gamma_1 = 0.292(0.055, 0.529)$, $\gamma_2 = -0.298(-0.994, 0.398)$; $R^2 = 0.913$. Thus, the three observations at $T < 12$ days <u>dominate</u> the parameter estimates of every mode. (Note that the 0.95 CL on γ_2 include zero.)

The models of the rival hypotheses, Power-law and LQ + time, are <u>non-nested</u> and thus the respective differences in residual sums of squared residuals cannot be used to discriminate between rival models of either dose-response or isoeffect, as would be the case for rival <u>nested</u> models. Nonetheless, there are some model discrimination criteria that are useful for both nested and non-nested rival models. For example, the correlation of observed and predicted levels of response, as measured by the multiple correlation coefficient R, or its square, R^2 can be used. (Gilchrist, 1984). Another is the Akaike Information Criterion (AIC) (Akaike, 1974). Still another is the posterior odds ratio (Leamer, 1978). We shall make use of the first and third of these model discrimination criteria.

It will be instructive to compare the rival models on the reduced (n=17) data set from which the outlying observation was deleted. The Power-law model is $D(\pi) = 10^{\alpha_0}N^{\alpha_1}T^{\alpha_2}$ where $10^{\alpha_0} = 18.294$, $\alpha_1 = \alpha_2 = 0.179$. The "fit" is assessed by both aggregate and case statistics. We have RSS = 100.90, $R^2 = 0.935$; the largest (absolute) residual, $|D-\hat{D}|$, is 4.81. The "LQ + time" model is $D(\pi) = (1 - \sqrt{\{1-(4\gamma_1/N)(\gamma_2 T + \gamma_0)\}}/(2\gamma_1/N)$, where $\bar{\gamma}_0 = -58.631$, $\bar{\gamma}_1 = 0.1465$, $\hat{\gamma}_2 = 0.680$. (Note that $\alpha/\beta = \gamma_1^{-1} = 6.826$ and $\gamma_2 = \gamma/\alpha$). We have RSS = 128.10, $R^2 = 0.919$ and the largest (absolute) residual is $|D-\hat{D}| = 6.03$. Obviously, the Power-law model is the more consistent with the Fowler 1991 data. From Bayes rule, the posterior odds ratio in favor of the alternative hypothesis H_1 (LQ + time) vs the null hypothesis H_0 (Power-law), given the Fowler 1991 data, S, can be written as (Leamer, 1978)

$$\frac{P(H_1|\,S)}{P(H_0|\,S)} = \frac{P(S|\,H_1)}{P(S|\,H_0)} * \frac{P(H_1)}{P(H_0)}$$

The first ratio on the right-hand side is the so-called Bayes factor, B. The second ratio is the prior odds ratio in favor of H_1. The data are said to favor H_1 relative to H_0 if the Bayes factor, B, exceeds one, that is, if the observed data S are more likely under hypothesis H_1 than under hypothesis H_0. For the above posterior odds ratio the Bayes factor can be expressed as $B = (RSS_0/RSS_1)^{n/2}n^{(k_0-k_1)/2}$, where RSS_j and k_j, j = 0,1 are, respectively, the residual sum of squares and the number of parameters in the regression model of the j^{th} hypothesis on a sample of size

n. We thus have B = $(100.90/128.10)^9$ = 0.117. Thus, in order to overcome the evidence of the sample data in favor of the Power-law model, the prior odds ratio for the LQ + time model must exceed 9:1. But, for the reasons given in section 7.5.1.1 above - "More is different" - this seems unlikely. The respective isoeffect surfaces for these two rival models are shown in Fig. 54a (Power-law) and 54b (LQ + time). The reduced (n=17) data set is superimposed on each plot. The respective residuals plots, $(D-\hat{D})$ vs $\hat{D}$, are presented in Fig. 55a and 55b. It appears from Fig. 55a and 55b that the "fit" of each model is quite good near the centroid of the data but markedly poorer near the extremes. However, a LOWESS smoothing of the absolute residuals suggests that this perception is somewhat misleading because of the increase in density of the points at both extremes (Cleveland, 1979). The fit is, in fact, only slightly better at the centroid than at the extremes - as would be expected. See Fig. 55c and Fig. 55d.

As we have shown in our analyses of the received (ED and ESD) models of the Mah et al data, there are implicit boundary conditions that must be satisfied by any model that is not simply a mathematical interpolation procedure, i.e., any model that is to describe a realizable process. One such boundary condition is that the model should not predict a realizable level of the response variable at unrealizable levels of the predictor variables, i.e., a positive response should not be predicted by negative levels of dose -or of time. It is evident from Fig. 54b that for the LQ + time isoeffect model levels of $D(\pi) > 0$ correspond to levels of $T < 0$ as well as $T > 0$. The former is nonsensical, suggesting that the LQ + time hypothesis is simply a mathematical interpolation, or smoothing, procedure for these data - and hence should be rejected since comparison of Fig. 55a and 55b discloses that it does not provide the better interpolation, or smoothing procedure as measured by the residuals. However, Fig. 54a discloses that the Power-law isoeffect model predicts that $D(\pi)$ vanishes for either $T < 0$ or $N < 0$ and hence is not invalidated by the implicit boundary conditions. (It will be recalled that both the ED model of Mah et al and the ESD model of Travis and Tucker predicted a positive level of response at a negative level of dose and hence could be rejected. Compare Fig. 54a and Fig. 54b with Fig. 42a and Fig. 42b. Also compare Fig. 54b with Fig. 45h. These are the isoeffect surfaces of the LQ + time models of clinical and experimental observations, respectively.)

N.B.: "We should also keep in mind that after a model is reported, the developer has little control over its use. A model may have been developed for interpolation purposes; however, once a model user finds that he has a good interpolation function, no amount of disclaimer by the model developer will keep the user from using the model to predict outside the region of the data and interpreting the coefficients to see how the different parts of the model work. This leads one to conclude that, in order for a model to be generally useful, it should have good extrapolation properties, and its coefficients should be acceptable estimates of the effects of individual terms." (R. Snee, 1977). And, "Furthermore, if this extrapolation performs poorly, it is almost always the model developer and not the model user who is blamed for the failure." (D. Montgomery and E. Peck, 1982). Note that the absence of readily accessible measures of the "leverage" of an observation (the hat matrix diagonals, h_i) for non-linear regression models such as the "LQ + time" isoeffect model makes it difficult to discriminate between those proposed treatment regimens (D,N,T) that are interpolations of the model and those that are extrapolations of the model - a crucial distinction for clinical use. For linear models, such as the linearized version of the Power-law model, however, a full range of regression diagnostics is readily available and the largest hat matrix diagonal, h_{max}, identifies the boundary of the so-called regressor variable hull (RVH). Points (N_i, T_i) within this ellipse are interpolation points. Points outside the RVH are extrapolation points. (Montgomery and Peck, 1982). Note also that the remarks of Snee serve to underscore the importance of the PRESS statistic and of the Ridge regression estimates that were examined above.

Although it may appear that the intrinsically non-linear isoeffect model of the LQ + time hypothesis that we have constructed is only slightly inferior to the contingently non-linear isoeffect model of the rival Power-law hypothesis, two remarks on model selection criteria provide an important context for this comparison. First, the values of R^2 for nonlinear regression models of such small samples (n=18) are notoriously - and optimistically - biased. The sample estimates of

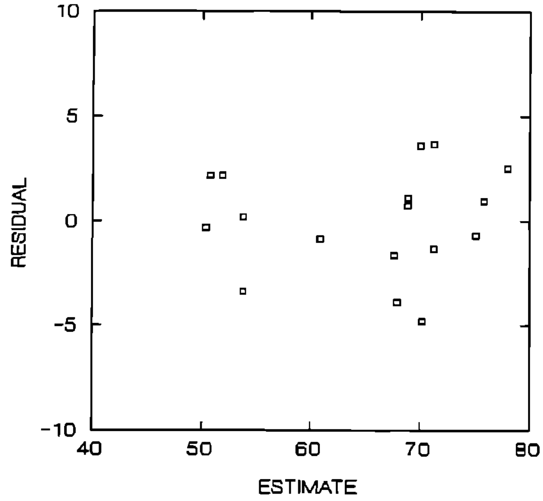


Fig. 55a. Scattergram of the residuals, $[D_i(\pi) - \hat{D}_i(\pi)]$ vs $\hat{D}_i(\pi)$, for the Power-law model of the reduced isoeffect data in Table 1 of Fowler 1991.

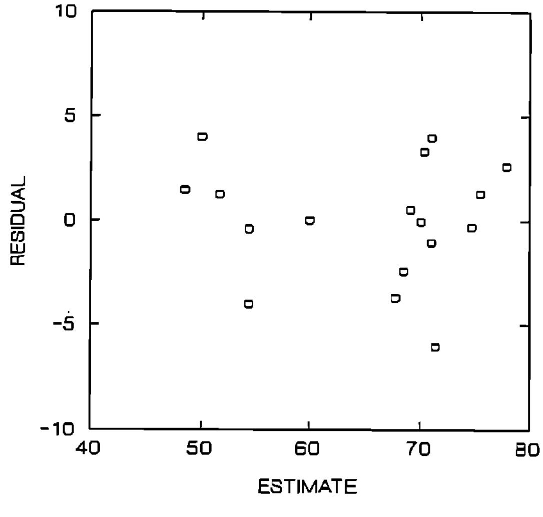


Fig. 55b. Scattergram of the residuals, $[D_i(\pi) - \hat{D}_i(\pi)]$ vs $\hat{D}_i(\pi)$, for the LQ + time model of the reduced isoeffect data in Table 1 of Fowler 1991.

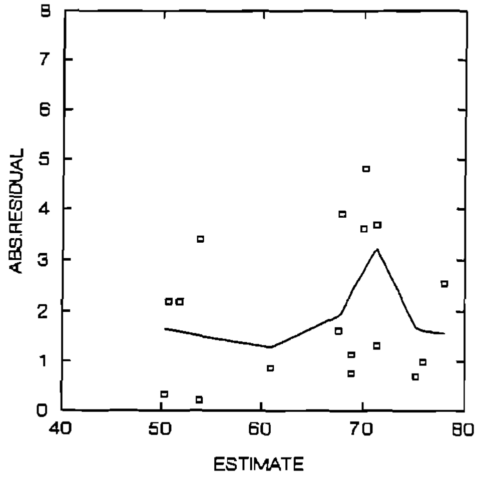


Fig. 55c. LOWESS smoothing of absolute residuals, $|D_i(x) - \hat{D}_i(x)|$, plot for the Power-law model of the reduced isoeffect data in Table 1 of Fowler 1991. The window is f = 0.50 (See Cleveland, 1979). The figure suggests that the "fit" is only slightly better at the centroid than at the extremes. Note that in general the "fit" of any regression model to any set of data will be better at the centroid than at the extremes.

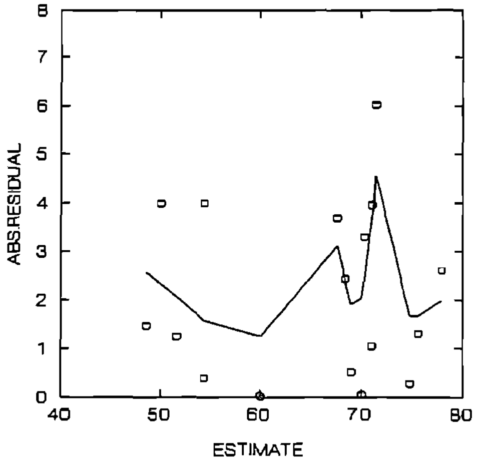


Fig. 55d. LOWESS smoothing of absolute residuals, $|D_i(\pi) - \hat{D}_i(x)|$, plot for the LQ + time model of the reduced isoeffect data in Table 1 of Fowler 1991. The window is f = 0.5 (See Cleveland, 1979). The figure suggests that the "fit" is only slightly better at the centroid than at the extremes. The figure also suggests that the Power-law model provides a better "fit" to the Fowler 1991 isoeffect data than does the LQ + time model.

model parameters, and functions thereof, are also biased for intrinsically nonlinear models of small samples. Second - and for both of these reasons - the contingently nonlinear isoeffect model of the Power-law hypothesis is to be preferred to the intrinsically nonlinear isoeffect model of the LQ + time hypothesis, for, as remarked above, "... in general, the model that comes closest to behaving as a linear model will be the preferred choice." (D. Ratkowsky, 1983).

17.3 Influential observations.

"The fact that a small subset of the data can have a disproportionate influence on the estimated parameters or predictions is of concern to users of regression analysis, for if this is the case, it is quite possible that the model-estimates are based primarily on the data subset rather than on the majority of the data."

D. Belsley, E. Kuh, and R. Welsch, 1980.

Finally, it will be extremely instructive to briefly examine one of the most recent studies of the putative effects of treatment duration on local control of tumor. This is reported in the paper by Keane et al 1992 in which, "The end-point in the analysis was ... control within the pelvis at 5 years for carcinoma of the cervix." In that paper, "A linear regression line was calculated for each tumor site and for T stage categories for carcinoma of the tonsil and by FIGO stage for cancer of the cervix. From the slope of the regression line the influence of treatment time on tumor control was expressed as a percentage decrease in control for each day of prolongation beyond the planned duration of treatment. The estimates of the slopes were tested against the null hypothesis of zero slope, using a t-test." T. Keane et al, 1992. However, it is immediately evident from Fig. 2 and 3 of that paper that for the cervical carcinoma data the estimates and inferences on the slopes are dominated by a single observation in each case. We have re-plotted Fig. 3A and Fig. 3B as Fig. 56a and Fig. 57a, respectively. The extreme observations are at T = 55 days in FIGO 1 and 2 (Fig. 3A) and at T = 72 days for FIGO stages 3 and 4 (Fig. 3B). If these observations are deleted the slope estimates change dramatically. Indeed, in the case of the set of n=10 observations for FIGO stages 1 and 2 (421 patients), the slope estimate for the sample with the right most observation deleted, does not differ significantly from zero. In the case of the n=7 observations for FIGO stages 3 and 4 (200 patients) the deletion of the observation at T=72 days reduces the slope from -1.23 to -0.90. Thus, the loss of tumor control with increase in overall time (of treatment) is not the 8.4%/week as has been reported (see Fowler 1993) but rather 6.3%/week at most. The respective influence plots are shown in Fig 56b and Fig. 57b in which the relative sizes of the symbols describe the relative influence of the several observations. Filled symbols identify observations with a negative influence on the slope and open symbols identify observations with a positive influence on the slope. These findings demonstrate that the published conclusions of Keane et al 1992 are not robust.

17.4 Conclusions.

"As is frequently the case, the rich diversity and vivid imagination of the theoretical literature is correlated with, and quite possibly a consequence of, a paucity of experimental evidence."

P. Rapp, 1986.

"A clinical trial cannot be adequately interpreted without information about the methods used in the design of the study and the analysis of the results."

R. DerSimonian et al, 1982.

"Much of what is published goes unchallenged, may be untrue, and probably nobody knows. Does anybody care? Do the methods used to obtain results matter anymore?

A. Neufeld, 1986.

"To be noteworthy a study may need to show that we have been wrong about something."

J. Marshall, 1990.

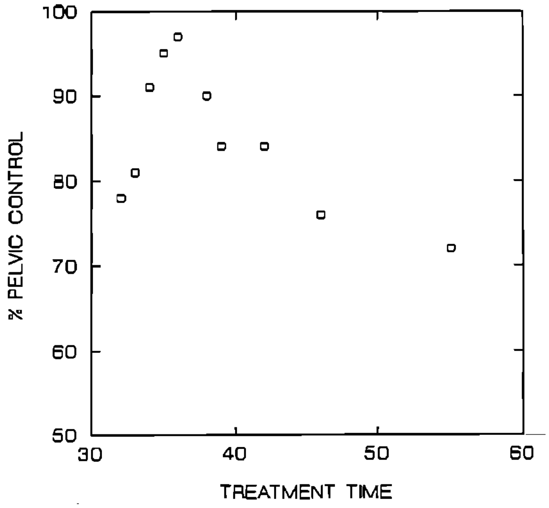


Fig. 56a. Scattergram of the data from Fig. 3A of the Keane et al 1992 report, (carcinoma of the cervix FIGO stages 1 and 2).

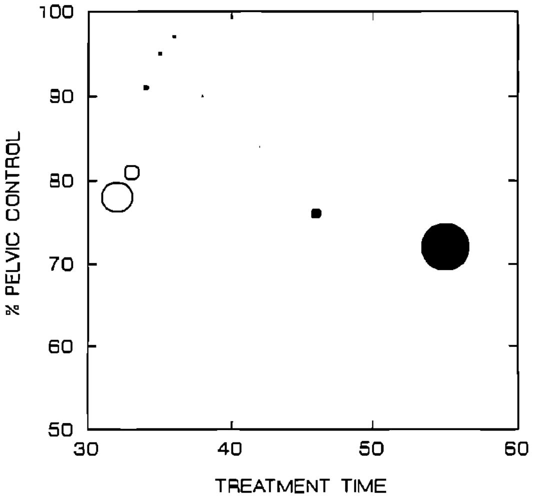


Fig. 56b. Scattergram of the row-deleted diagnostic, R - R(i), $1 \leq i \leq n = 10$, where R is the Pearson correlation coefficient for the data of Fig. 56a. The size of each symbol is proportional to the decrement, R - R(i), for the i^{th} observation where R(i) is the correlation coefficient for the data with the i^{th} observation deleted. The symbols are as in Fig. 45f.

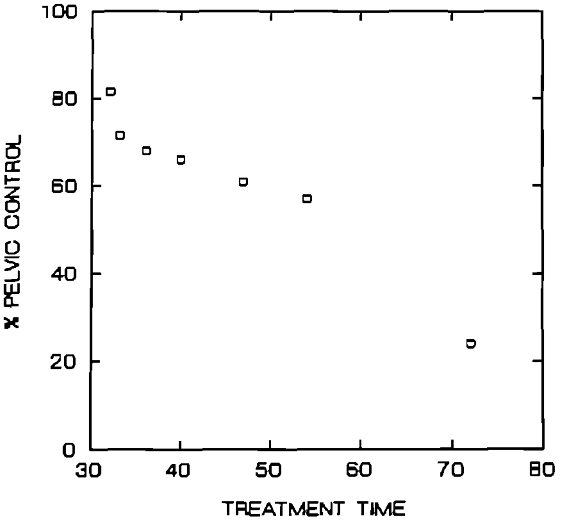


Fig. 57a. Scattergram of the data from Fig. 3B of the Keane et al 1991 report (carcinoma of the cervix FIGO stages 3 and 4).

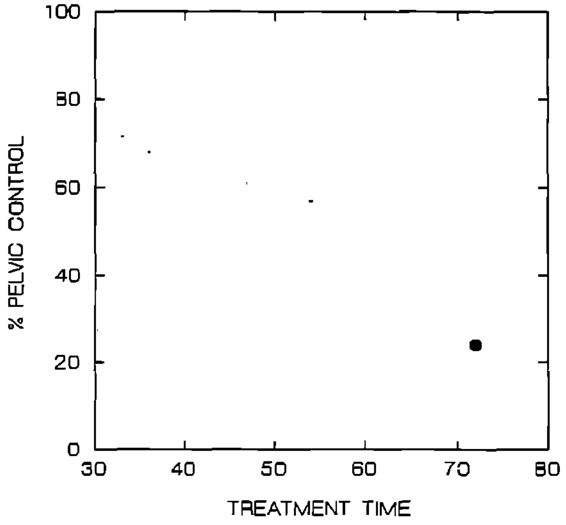


Fig. 57b. Scattergram of the row-deleted diagnostic, R - R(i), $1 \leq i \leq n = 10$, where R is the Pearson correlation coefficient for the data of Fig. 24a. The size of each symbol is proportional to the decrement, R - R(i), for the i^{th} observation where R(i) is the correlation coefficient for the data with the i^{th} observation deleted. The symbols are as in Fig. 45f.

In assessing the validity of a reported finding it is most important to evaluate for statistical adequacy the methodology by which the finding is demonstrated. Here a homely analogy may be of use: The best - indeed the only - evidence for the validity of the randomness of a random sample is the evidence that the sample was generated by a demonstrably "random process". In the same way, the best evidence for the validity of a scientific finding is the evidence that it was obtained by statistically adequate methods of design, analysis, inference, etc. Thus, the methods by which Withers et al 1988 supported their finding of a two-phase isoeffect curve in the data of their Figure 1 are rather sui generis and ad hockery and therefore their finding is dubious. Hence, the statistically adequate methods we have demonstrated - non-parametric and non-linear regressions - greatly strengthen one of their conclusions, namely, that a linear spline with "knot" at $T \simeq 25$ days is not inconsistent with the isoeffect data of Fig. 47. (However, we found that neither of the rival logit models of the dose-response data "fit the data," suggesting that there is much more variation in the observed response than can be captured by any model based on the information available in their Table 1. Since the sample estimates of the parameters and functions thereof will be biased for a model that does not fit the sample data, this finding casts doubt upon the validity of all of their inferences on the dose-response relationship, including, especially, the validity of their accelerated repopulation hypothesis that is inferred from their Fig. 1.) Similarly, the report of Fowler 1991 that an LQ + time model fit his isoeffect data is also based upon ad hockery and sui generis methods. However, we demonstrated that the appropriate nonlinear regression model of the LQ + time hypothesis was not inconsistent with these isoeffect data and, moreover, gave parameter estimates that were not inconsistent with a priori information. (See above remarks by Robins and Greenland, 1986.) Nonetheless, we also demonstrated that the rival Power-law hypothesis gives the isoeffect model of choice for these data, based on both statistically adequate criteria of fit and the posterior odds ratio - as well as on the inherent superiority of linear models over nonlinear models described by Ratkowsky 1983.

We have now evaluated (in sections 16 and 17), by means of secondary analysis, three published (Mah et al, 1987; Withers et al, 1988; Fowler, 1991) - and widely celebrated - studies of the role of the "time factor" in modulating the level of binary responses in normal and tumor tissues exposed to fractionated doses of ionizing radiation. We have also evaluated one of the most recent of such studies. A role for the time factor in clinical radiation response is a durable problem. As Barton et al 1992 have remarked: "The relationship of the overall treatment time and total dose to the outcome of a course of irradiation remains controversial more than 40 years after the first major attempt at a theoretical description."

None of the three earlier studies included data that were sufficiently strong to provide an empirical basis for discriminating unambiguously between the models of the two rival hypotheses: LQ + time and Power-law. The Withers et al 1988 (dose-response) data rejected both of the rival dose-response models on the evidence of both aggregate and case statistics. Both of the final rival models included as covariates the anatomical site and the stage of disease, as well as the treatment variables, dose, time, and fractions or dose/fraction. The estimates of the coefficients of each covariate were much greater than their respective standard errors for both models save for the term D^2/N in the LQ + time model and logN in the Power-law model, neither of which were statistically significant. For the LQ + time model the estimates of the ratios of parameter estimates were $\alpha/\beta = -105.92$ and $\gamma/\alpha = -0.551$. There were no other covariates included in the Withers et al data (59 studies) that might account for this lack of fit - other than date of study, for which the coefficient in the logit model was not significant.

The Fowler 1991 (isoeffect) data rejected neither isoeffect model. But, on statistically adequate criteria for discriminating between non-nested rival models (R^2, posterior odds ratio, and PRESS) the Power-law isoeffect model provided a much better representation of these data than did the version of the LQ + time isoeffect model that was reported in the Fowler paper.

However, we also constructed a nonlinear isoeffect model of the LQ + time hypothesis on these data, which "fit" much better than did the Fowler isoeffect model. But, because of the characteristic weaknesses of nonlinear models compared with linear models having the same number of parameters and approximately equal concordance with the data - R^2 = 0.923 (Power-law) vs R^2 = 0.913 (LQ + time) - the Power-law model is still the model of choice on this criterion - as well as on the Bayesian criterion of the posterior odds ratio.

It should be noted that Power-law isoeffect models of both the Withers et al data and the Fowler data were not inconsistent with the old NSD model: $D(\pi) = 18.00N^{0.24}T^{0.11}$. Thus, although it might seem possible, a priori, to retrieve some information on the LQ + time models of dose-response of tumors from the Withers et al data (the object of that study) such a possibility would surely seem to be foreclosed from the beginning for the LQ + time isoeffect models of normal tissue "tolerance" on the Fowler data (the object of that study).

The clinical pneumonitis (dose-response) data of Mah et al 1987 are by far the strongest data of these three studies. (For one thing there is only within-study variation; there is no between-study variation, as was the case of the Withers et al data.) The Mah et al data did not reject either of the two rival models. However, the coefficient of the dose term in the Power-law model exceeds its standard error by a factor of two whereas the coefficient of the dose term in the rival LQ + time model is less than its standard error by nearly a factor of ten. The observed lack of significance in the fractions and times terms in the Power-law model can be rather confidently ascribed to the high degree of multicollinearity in these data for this model - i.e., a model-sample interaction. A similar explanation of the observed lack of significance of the fractions and time terms in the LQ + time model cannot be supported by any feature of the data. This comparison of rival models on these data suggests that the LQ + time model is mis-specified.

Thus, our findings for this section may be summarized in the following way: First, although the data presented in the three earlier published studies has many weaknesses, it provides stronger support for the Power-law hypothesis than for the LQ + time hypothesis on the evidence of our secondary analyses. (Thames (1988) has observed that, "Therefore, no matter how sophisticated the fitting techniques or careful the design of the experiment, no power-law representation of proliferative response can be expected to describe the data adequately." This suggests that our findings are "noteworthy" - on the Marshall 1990 criterion quoted above.) Second, the conclusions, estimates, and inferences of Keane et al 1992 on the effect of treatment duration on local control of carcinoma of the cervix are not robust since their validity depends strongly on a single extreme observation of the sample (T = 55 days for FIGO 1 and 2; T = 72 days for FIGO 3 and 4).

The failure to validate by secondary analysis the published findings of the three earlier and oft-cited studies (Mah et al, 1987; Withers et al, 1988; Fowler, 1991) on the role and functional form for a time factor in the radiation responses of both normal and tumor tissues and thus to come to closure on the question is one sure sign that the statistical methodology deployed in the published primary analyses had several weaknesses. (Comparison of Fig. 42a with Fig. 42b and Fig. 45f with Fig. 52b and Fig. 56b and Fig. 57b discloses that the same mistake was made repeatedly. But, "When a paper containing incorrect results (not necessarily through statistical mistakes) is published there may be serious consequences although surprisingly this does not seem to be generally appreciated: ... (v) If the results go unchallenged the researcher(s) involved may use the same substandard statistical methods again in subsequent work, and others may copy them. ..." D. Altman, 1982. The above findings that the reported "straightness" of several data sets is not robust can be aptly summarized, with only a slight change in the original, by the well-known lines from Shakespeare's Twelfth Night, Act II, Scene 5: "Some data are born straight, some achieve straightness, and some have straightness thrust upon them!") The studies by Withers et al 1988, Fowler 1991, and Keane et al 1992 disclose once again how "strenuous and devoted" are the attempts of many investigators to "force nature" into the box, y = a + bx that was provided by their professional education (see section 7.5.3.3 above). It also suggests that debate on the questions of the role and functional form for a time factor has been prematurely foreclosed among a significant group of investigators. It finally suggests that the data required to determine the role of a time factor - and the functional form of that factor - remain to be collected. At present the role and the currently proposed forms (save perhaps for sign) are largely matters for conjecture. The first step in bringing the issue to an acceptable closure is to collect appropriate data by well-established statistically adequate methods.

The first desideratum is obviously that the sample be of a "statistically adequate" size: It is important to use large samples. Only then can one be reasonably sure that results obtained can be replicated on the basis of different samples drawn from the same population. "... a useful rule to observe is that the sample size in multiple regression problems should be at least 100 or at least 20 times the number of [predictor] variables, whichever is larger." (Lindeman et al, 1980) It is apparent that none of the studies we have examined in this report meets this minimal criterion. (As we have remarked before, it often seems that many investigators fail to take data very seriously - since they take so little of it.)

Viewed from the perspective provided by the findings of these and of our previous secondary analyses, the current persistence of the LQ and "LQ + time" models should perhaps be taken less as evidence for their validity, than as providing yet another instance of what Nisbett and Ross (1980) have called "belief perseverance," an epiphenomenon that was first - and most elegantly - described by Francis Bacon in 1620: "The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that ... the authority of its former conclusion may remain inviolate." Or, more recently, and less elegantly, Nisbett and Ross observe that, "Conflicting evidence is treated as if it were supportive of beliefs, impressions formed on the basis of early evidence survive exposure to inconsistent evidence presented later, and beliefs survive the total discrediting of their evidence base."

Footnotes

[1] We have defined those beliefs, estimates, inferences and models as received if they appeared to us to represent the consensual preferences of the peer-group (e.g., received standard English is the dialect spoken in the British public schools. It is sometimes referred to as RP - Received Pronunciation).

The present report demonstrates that these received beliefs often arise in epistemological weaknesses residing in statistical solecisms that were committed in the acquisition and analysis of biological data and which either suggest - or entail - or otherwise support fallacious arguments leading to false conclusions and thence to a false ontology. (p. 2)

[2] As is the case with most astringent aphorisms this, too, had a history: Cicero delivered himself of the same observation about eighteen hundred years earlier. It also had a future: Similar observations have been made recently concerning the performance of the adversarial expert witness in toxic tort cases. See part 13 of this report. (p. 17)

[3] These views of Ziman and Kuhn, as well as many of those of Jeffreys and Popper cited elsewhere in this report were anticipated by C.S. Peirce, an American physicist/philosopher of the nineteenth century who labored under several burdens not the least of which was a very infelicitous style of expression. That he could anticipate - and to some degree reconcile - such disparate views as those of the more articulate Kuhn and Popper who followed him is evidence of his extraordinary capacities. He is widely regarded as America's foremost philosopher and polymath - a "multifaceted genius."

We have depended heavily on the views of Whewell, Popper, Kuhn, and Peirce anent the history and philosophy of scientific enquiry for several reasons. First, of course, Popper and Kuhn appear to be the two philosophers that are most widely quoted by practicing scientists (See for instance, Buck, 1975; Jacobsen, 1976; and Susser, 1977). Second, Kuhn and Peirce were trained as physicists and therefore their views may be expected to be most likely to be found "legitimate" by other physicists. William Whewell, was an early nineteenth century Cambridge polymath (mathematician, philosopher, mineralogist, classics scholar) and word-smith to Michael Faraday (Whewell coined the terms "anode", "cathode", "scientist", and "physicist".) Finally, it has appeared to us, that Thomas Kuhn has, so to speak, "written the script" for the way biomedical science is practiced in the last quarter of the twentieth century. (p. 37)

[4] Recently R.C. Bolles (1988) has presented a more ecumenical version of Descartes' principles: "Rule 6

The sixth and perhaps most enlightened principle I wish to convey, the sixth method for avoiding statistics, is that you should talk to Buddha, or somebody like that."

...

"And when I plotted this against that, I got an ugly curve, a curve that bent this way and that. I remember thinking to myself, as I sat staring at these ugly data, that Buddha would not have liked them. They were not pretty, they did not harmonize, they did not look like a part of nature. They made sense in that the curve moved upward in the direction one would expect, but the curve was ugly. This was not the way Buddha would have arranged things. So I fussed with the data."

...

"I suddenly understood something of the spiritual aspect of science. To my amazement and delight, I saw I had a straight line. When I saw that, I knew that I had at last discovered something special that Buddha, or someone like that, had put into nature hoping that someone like me would discover it. Here was a real law of science.

When you have something that one of the gods wants you to discover, you certainly do not have to do a statistical test. No one has to prove the linearity of the data. There it is. By god, it is linear. With such a discovery in hand, statistical tests become irrelevant. And irreverent.

...

"When you refuse to let the statisticians close the doors on you, then you are free - free to discover the great world that lies out there." (p. 39)

[5] One such goal may be publication in which it is frequently more important (to a scientific career) that views be held in common than that they be correct. As P. Woolf (1988) remarked, "If you tell me how the score will be kept, I'll tell you how the game will be played." (p. 44)

[6] Since the ϵ_i describe, "... the limits and validity of our knowledge" - obtained from the data - that is represented by μ_i, it is an apt locution. (p. 50)

[7] The physicist Sir William Hamilton has expressed the canon of the Franciscan, William of Ockham, as, "Neither more, nor more onerous, causes are to be assumed than are necessary to account for the phenomena." Ockham's expression is still more parsimonious: "What can be done with fewer is done in vain with more." If, as seems to be the case, the lawyer Francis Bacon is accepted as the father of modern empirical and experimental science (Bacon coined the terms "experiment" and "experimental philosophy") then the two skeptical Franciscan friars, Roger Bacon in the thirteenth century and William of Ockham in the fourteenth, must be regarded as the grandfathers of that enterprise. (p. 58)

[8] The statistician Karl Pearson, like the ecclesiastics Osiander and Berkley and the physicists Duhem and Mach before him, held the (logical positivist) view of the nature and role of scientific laws: "The law ... is a brief description of how ... . It does not tell us why ... . It simply resumes, in a few brief words, the relationships observed between a vast range of phenomena. It economizes thought by stating in conceptual shorthand that routine of our perceptions ..." (Pearson, 1892/1922). Or, as Mach himself would have it: "Physics is experience, arranged in economical order." (p. 58)

[9] The prejudice against statistics in Science has an honorable pedigree that can be traced back to at least the early 19th century. The prejudice was given institutional expression at the beginning of the 20th century by the British Royal Society: In 1901, the Council of the Society, in response to a paper containing statistics in application to a biological problem that was submitted by Karl Pearson, passed a resolution, "requesting that in future papers mathematics should be kept apart from biological applications."

Our reviews of the current radiobiological literature, described in Annexes I-IV of this report, have disclosed that the writ of the Royal Society runs wider - and longer - than one might have at first supposed: Current radiobiological praxis has achieved - and maintained - what the Council would surely regard as a seemly separation of mathematics from biological applications. (p. 178)

[10] In 1962 L. Wollin attempted to obtain their raw data from several scientists for secondary analysis. Of the 37 to whom requests were made, 32 responded. More than two thirds of the latter reported that their raw data had been inadvertently lost or destroyed. Of the 7 sets of data that Wollin was able to re-analyze, 3 disclosed that gross miscalculations had been made by the original investigators that would have required substantial changes in the conclusions originally published. (p. 199)

[11] "They [the Bourbons] have learned nothing and forgotten nothing." (p. 200)

C. Talleyrand

[12] As is characteristic of many of the so-called Baconian, or empirical, sciences (chemistry, metallurgy, thermodynamics, etc.), radiation biology derives largely from a craft - radiation oncology. (p. 204)

[13] However, it should be recalled that in general scientific practice, "... there are seldom many areas in which a scientific theory, particularly if it is cast in a predominately mathematical form, can be directly compared with nature" ... "Furthermore, even in those areas where application is possible it often demands theoretical and instrumental approximations that severely limit the agreement to be expected." (T. Kuhn, 1970a)

Max Born (1949) has also remarked on the exiguous number of observed instances upon which a natural law is based: "... no observation or experiment, however extended, can give more than a finite number of repetitions"; thus, "the statement of a law - B depends on A - always transcends experience. Yet this kind of statement is made everywhere and all the time, and sometimes from scanty material". (emphasis added) (p. 205)

[14] It should be noted that the complete van der Kogel data on rat hind leg paresis have never been published. It should also be noted that three of the data sets on which the risk estimates presented in the BEIR III report are based are not presented in that report. The breast cancer incidence data were published in another paper before that report was issued. The leukemia incidence data were published in another paper after that report was issued. The non-leukemia cancer mortality data have never been published. (p. 205)

[15] The importance of the role of criticism to the growth of knowledge, as described above can be further emphasized by an anecdote of the chemist Linus Pauling, who upon being asked by a student as to how to go about having good ideas replied, "You have a lot of ideas and you throw away the bad ones." Statistical methods - especially regression diagnostics - can assist in identifying, "the bad ones". Note that Pauling's method is quite consistent with Popper's method of "conjecture and refutation" that he has proposed as the route by which scientific knowledge accumulates. See also the (1964) essay "Strong Inference" by the chemist John Platt. (p. 212)

[16] "Toxic tort cases may be loosely defined as those in which plaintiffs seek compensation for harm allegedly caused by exposure to a substance that increases the risk of contracting a serious disease, but generally involve a period of latency or incubation prior to the onset of this disease. Exposure to radiation and the use of pharmaceutical drugs or products fall within this loose rubic." (Black, 1989). (p. 213)

[17] "The South Sea cargo cults provide a clear example of magical thinking. During World War II airplanes landed and unloaded food and materials. To bring this about again, the natives built fires along the sides of makeshift runways. They have a ceremonial controller who sits in a hut wearing a wooden helmet complete with bamboo bar antennas (sic). They recreate the pattern observed in the past and wait for the planes to land. No airplanes land, yet the ritual continues ... several studies of human learning indicate that such behaviour is common in our own activities." P. Diaconis, 1985. Thus, many investigators now repeat, as a ritual performance, mathematical maneuvers - say construction of a "straight-line" graph - that hitherto, in other circumstances, have yielded a desired result: valid information on some inherent feature of the process generating the observations. But in several of the current contexts in which these rituals are now performed they only serve to misinform the practitioner - no airplanes are landing. (p. 234)

[18] It was Aristotle who first recommended that, with their hypotheses, scientists ought to "account for the observed facts" rather than "do violence to them by trying to squeeze or to fit them into their theories" (See Popper, K. Conjectures and Refutations, 1965b). (p. 264)

Appendix I. The Effect of Volume of Irradiated Tissues on Radiation Response.

"In general, it's a good thing to know what you are doing and why you are doing it."

R. Merton, 1975

"What Do We Believe?"

Since scientific knowledge is consensual, or group, knowledge (Ziman, 1968), it is important to establish that any studies selected for evaluation are generally accepted by the peer-group. Of course, if a paper is published in the peer-reviewed literature and is frequently cited in that literature, that is, as they say, prima facie evidence that it is representative of received opinion. However, it is still stronger evidence of the acceptability of a study to the peer-group if its findings are described and perhaps further elaborated - or even only mentioned - in a published and authoritative review of the topic with which it is concerned. For example, Fowler's 1984 review of dose-response models and Cohen's 1982 review of volume effects models are authoritative reviews. The NCRP 64 (1980) review of dose-time models is another example. The BEIR III (1980) report is yet another.

In 1988, an authoritative review, "Treatment Volume and Tissue Tolerance", by Withers, et al notes that, "The effect of changes in treatment volume on tolerance dose has been an area of uncertainty and debate among radiation oncologists, not only regarding its mechanism and quantitation, but about its very existence. Some reports claim an effect ... whereas others do not." Several reports are cited as providing evidence for each point of view but there are a much larger number of papers claiming existence of a volume effect than denying it (20 to 5) and among those studies cited as claiming an effect are an earlier review by Cohen (1982) and two original papers by von Essen (1960, 1963). Cohen's review and the latter two papers have been cited 9, 41 and 21 times, respectively, between the date of first publication and October 1989.

It will be recalled that systematic review of several research studies in order to come to a general conclusion as to which of several rival hypotheses seem most supported by the preponderance of the empirical evidence in the literature now falls under the rubric of meta-analysis. The simplest meta-analysis is simply to "count the votes", as could in the Withers et al, 1988 review. Hedges and Olkin (1985) remark that, "The inference procedures used in conventional vote-counting procedures are inherently flawed and likely to be misleading." The most complex is the hierarchical Bayesian methods described by DuMouchel and Harris (1983). Mixed estimation (See section 7.2.4 and Annexes II and IV on Mixed estimation and section 14.3 on Bayesian hierarchical meta-analysis) may be usefully regarded as a meta-analysis of intermediate complexity. As DuMouchel has remarked, "Critical judgments about the accuracy, quality and/or methodologies of the individual studies are necessary in order to combine them intelligently." Such critical judgments are usually much assisted by secondary analyses of the individual studies before seeking to combine the information therein by any method. Even as in the present case. In meta-analysis one should always re-analyze - or at least closely re-read - the reports of the studies that are to be integrated. That this is desirable in even the simplest form of meta-analysis - the "vote-count" - is now illustrated. (The moral seems to be that nothing can be taken for granted in this business.)

Cohen, in his 1982 review, "The Tissue Volume Factor in Radiotherapy," remarks that, "Field size or irradiated volume is one of the four important variates affecting the severity of the reaction to irradiation (the other three are dosage, fraction number, and overall time). Inclusion of volume factors in existing formulae for dose-time adjustment would be a significant advance in standardizing treatment schemes in clinical practice," and then provides a concise account of the received wisdom on the matter. Note that the evidence for a volume effect appears to be unequivocal - apodictic - in Cohen's account: "It is generally true that with normal tissues, the dosage associated with a given risk of radiation injury is smaller than when the treated volume is large."

"With tumors, the dose-volume relationship is the reverse of that encountered with normal tissues. Large tumors require higher doses to achieve the same probability of local control." Cohen then refers to the 1960 von Essen paper which he summarizes as follows: "In his analysis of factors determining success or failure in radiotherapy of cancers of the skin and lip, von Essen derived best-fitting exponents in the empirical power function for tumor volume and skin area. The exponents were shown to differ in both magnitude and sign in the two situations. Since normal tissue tolerance is inversely related to field size the exponent for normal tissue has a negative sign. It's magnitude was estimated to be y = -0.27. Since the effective tumor dose varies directly with tumor size, the exponent for tumors must be positive, and the best estimate for this parameter was given by y = +0.14. Von Essen incorporated these parameters into isoeffect formulae for tumor control and normal tissue tolerance, and developed a complex iso-response grid as a guide to management." Cohen's is an authoritative review and hence can be taken to fairly represent the received wisdom on the matter, as least in the main. However, it appears that Cohen misread von Essen's 1960 paper. We shall discuss this more fully below, but it is not an unusual occurrence. As Derek Price has remarked, "Researchers have a strong urge to write papers but only a relatively mild one to read them."

Why Do We Believe It?

It appears from von Essen's paper that the most cogent reason for believing the evidence presented for the existence of a "volume effect" are his reports of statistically significant values of the Pearson chi-squared statistic, $\chi^2(1)$, for the several cross-classifications of his data:

a. Small Tumors (< 10 cm^2)
$D = 2680T^{0.14}$ (Ablation)
$\chi^2(1) = 14.2$, $P < 0.01$

b. Large Tumors (> 10 cm^2)
$D = 2950T^{0.14}$ (Ablation)
$\chi^2(1) = 0.37$, $P > 0.50$

c. Small Fields (< 10 cm^2)
$D = 1950T^{0.27}$ (Necrosis)
$\chi^2(1) = 21$, $P < 0.01$

d. Large Fields (> 10 cm^2)
$D = 17001^{0.27}$ (Necrosis)
$\chi^2(1) = 6.2$, $P < 0.01$

von Essen further summarizes his evidence as follows: "The data derived from this material suggest that the isoeffect line for 99 percent probability of cure for the initial treatment of small tumors have a slope of 0.14 with an intercept at the single treatment axis of 2680r. If it is assumed that larger lesions have parallel isoeffect slopes with progressively higher intercepts at the single treatment axis, then a family of parabolas, transformed to lines on logarithmic coordinates, will have the general equation $D = kT^{0.14}$.

Similarly, a family of parabolas can be developed for a certain probability of skin necrosis. In the data presented, a slope of 0.27 was derived for the line limiting the probability of skin necrosis to about 3% for smaller radiation fields. If the isoeffect lines are parallel for different field sizes, then the family of parabolas will have the equation $D = kT^{0.27}$."

...

"Super-imposition of these two sets of parallel lines with different slopes illustrates the proposed model. ... only two of the isoeffect lines have been derived directly from the data. The construction of all the other curves parallel to these two originally derived curves is, then, hypothetical" (Emphasis added).

Judging from von Essen's description of the respective circumstances of a)-d), the two "originally derived" isoeffect curves referred to are for Small Tumors and Small Fields.

Should We Believe It?

"It is apparent that the crucial elements in risk analysis are: (a) the numbers; and (b) the methodology. It is on them that critical scrutiny must concentrate."

I. Hoos, 1980

The von Essen dose-response data are binary: Necrosis/No Necrosis (skin) and Ablation/Recurrence (skin cancer) at each level of treatment. Each level of treatment is specified by the covariates total dose, D (Roentgens) and total duration of treatment T (days). Thus, the appropriate model is that for a Binomial distribution of response, either probit or logit: $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 is the dose and x_2 is the time (or the respective logarithms). It will be convenient for the moment to consider only the probit model. Then $z = \Phi^{-1}(\pi)$ and $x_1 = \log D$, $x_2 = \log T$ where $0 \leq \pi \leq 1$ is the proportion of responders at $\underline{x}^T = (x_1, x_2)$, $\Phi(\ )$ is the Normal distribution function. Then the dose-response model may be written as, $z = \underline{x}^T \underline{\beta}$ where $\underline{x}^T$ has been augmented by $x_0 = x^0 = 1$ and z is the probit transform. The goodness-of-fit of the model and data should be assessed by examining both aggregate (chi-squared) and case (chi-residuals) statistics. The values of the ratios $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$, $j = 1, 2$, can be used to identify those covariates that are significantly associated with the response; typically, $\hat{\beta}_j\sqrt{Var(\hat{\beta}_j)} > 2$ identifies a statistically significant association.

However, it is important to note that this normative procedure was <u>not</u> the method followed by von Essen. Instead, in his procedure it was simply assumed <u>a priori</u>, on the basis of a previous paper by Strandqvist that this dose-response model was valid for these data, or, more precisely, that the cognate isoeffect model that is entailed therein, was valid; neither the goodness-of-fit of the model nor the significance of the parameter estimates was examined. The <u>isoeffect</u> model that corresponds to the above dose-response model, $z = \underline{x}^T\underline{\beta}$, is $x_1(\pi) = (z - \beta_0)/\beta_1 - (\beta_2/\beta_1)x_2$. However, in the von Essen paper, it is simply written as $D(\pi) = D = kT^n$ or $\log D = \log k + n \log T$. For several different strata identified by ranges of volumes, V, or areas, A, of irradiated tissues, von Essen attempts to estimate π, k and T from his data by constructing, within each stratum, a cross-classification in which one dichotomy, the binary response, is given a priori and the other dichotomy, defined by the isoeffect curve, is determined from the sample data. von Essen then represents by a graph in which the respective families of isoeffect curves for skin and skin cancer are super-imposed, a hypothetical relation between V (or A) and k for skin and for skin cancer. These families of isoeffect curves had the following forms: $D_1 = k_1T^{0.14}$ (tumor). $D_2 = k_2T^{0.27}$ (skin), where $k_j = k_j(A)$, $j = 1,2$. This particular graph has proved to be a rich source of diverse estimates of the exponent of the volume, or area, factor in power-law forms of these relations. Two are given in Table <u>1</u>.

von Essen has presented the major part of the evidence for his subsequent estimates of the form and parameters of his model of the dependence of the binary response of irradiated tissues upon the area (volume) of the irradiated tissues in the form of tests of the null hypothesis of independence of two dichotomies by which his data can be classified. One of the dichotomies is the <u>response of a tissue</u> irradiated at treatment level (D_i, T_i): Ablation/Recurrence of tumor or Necrosis/No Necrosis of skin. The second dichotomy is the relative position of the treatment level (D_i, T_i) with respect to a given line: "above the line"/"below the line."

Although the evidence was not presented in such a format in the von Essen paper, such evidence is more clearly - and more commonly - presented as a fourfold (2x2) contingency table. It is important to recall some of the essential features of such tables. As an example, Table <u>2</u> presents the results of classifying a sample of size n = a+b+c+d in two different ways according to 1) whether they do (A) or do not ($\bar{A}$) possess the attribute A, and 2) whether they do (B) or do not ($\bar{B}$) possess the attribute B. Note that the inferences on the independence of the two dichotomies that are based on such tables require that the category structure of both of the dichotomies, $A/\bar{A}$ and $B/\bar{B}$, are completely known and specified, <u>a priori</u>; that is, their specification must be independent of any sample information. If this is the case, then the statistic, $\chi^2 = n(ad-bc)^2/[(a+b)(c+d)(a+c)(b+d)]$, is distributed approximately as chi-squared with one degree of freedom on the null hypothesis that the two attributes are <u>independent</u>. (The approximation can be improved by introducing the so-called Yate's correction, however there is currently some question as to whether the correction is really desirable. See S. Weinberg (1983).) If $\chi^2 > 3.84$ then the hypothesis of no association (independence) is <u>rejected</u> ($p \geq 0.05$).

However, the cross-classifications of his data that were established by von Essen do <u>not</u> satisfy the assumptions required for the validity of the chi-squared test of hypothesis that he employed. In von Essen's cross-classification one of the two dichotomies in each of the tables ("above the line"/"below the line") is achieved by partitioning a continuous variable. And although the category structure of the binary tissue response attribute in each table can be completely specified a priori (Ablation/Recurrence and Necrosis/No Necrosis) the structure of the other category (the dichotomy, "above the line"/"below the line") cannot be precisely specified independently of sample information on the intercept and slope of "the line": "A simplified derivation of isoeffect curves by empirical fitting of smooth curves relating a certain probability

Table 1. Estimates of Exponents of A in Empirical Power-Law Descriptions of "Volume Effect" in Skin and Tumor. von Essen 1960 data[#]

Exponent		
Skin	Tumor	Source
-0.16	0.16	von Essen 1973
-0.16	0.075	von Essen 1963

[#] von Essen: $\bar{N}/N$ and $\bar{A}/A$.
565 Lesions including 45 A (8%) and 29 N(5%)
Small Fields/Tumors (< 10 cm^2)
Large Fields/Tumors (> 10 cm^2). Includes "100 cm^2 Tumors"(?)

Table 2. The Four-fold (2x2) Contingency Table.

		Predicted		
		B	$\bar{B}$	
Observed	A	a	b	(a + b)
	$\bar{A}$	c	d	(c + d)
		(a + c)	(b + d)	n

Chi-squared on 1 df (without Yate's correction):
$\chi_u^2(1) = n(ad - bc)^2/[(a + b)(c + b)(c + d)(a + c)]$
Pearson's Contingency Coefficient:
$\phi = \sqrt{}(\chi_u^2/n)$. $-1 \leq \phi \leq 1$.
Odds Ratio:
$OR = ad/bc \qquad 0 \leq OR < \infty$
Standard Error of OR:
$se(OR) = OR(a^{-1} + b^{-1} + c^{-1} + d^{-1})^{1/2}$
For cells that are either empty or have small numbers, we have (Fleiss, 1973)
$OR^{\#} = (a + 0.5)(d + 0.5)/(c + 0.5)(b + 0.5)$
$se(OR^{\#}) = OR^{\#}[(a + 0.5)^{-1} + (b + 0.5)^{-1} + (c + 0.5)^{-1} + (d + 0.5)^{-1}]^{1/2}$

Note that for a logistic regression model of binary response on a dichotomous (0, 1) risk factor, x, $\log[\pi/(1-\pi] = \beta_0 + \beta_1 x$, we have $\log(OR) = \hat{\beta}_1$ and $se\{\log(OR)\} = [a^{-1} + b^{-1} + c^{-1} + d^{-1}]^{1/2} = \sqrt{}(Var(\hat{\beta}_1))$.

of cure to the coordinates of dose and time is attempted. The effectiveness of the curves to separate cured and non-cured lesions is tested by Chi-squared analysis" (von Essen, 1960). The fundamental problem is that of constructing and evaluating a discriminant function that will optimally distinguish, or "separate", the complementary classes of tissue response, given the information (D_i, T_i) on their treatment.

Although von Essen was surely unaware of it, he seems to have estimated the slope and intercept of each of the four "isoeffect" curves to which the greater part of his data were assimilated, by a procedure that was later described by M.G. Kendall (1966) in a quite different context as a "distribution-free discriminant analysis"; an iterative clustering procedure, based on the order statistics of the sample, in which the investigator begins with an initial partition of a two-dimensional scattergram of a bivariate sample defined by the position and orientation of "the line" and tries to make the two partitions (tissue response/non-response and above/below "the line") more consistent by subsequently reassigning tissue responses from one category to another, e.g., from "above the line" to "below the line" while adjusting the position and orientation of "the line". As evidence that this was the procedure actually followed by von Essen we cite the following remark: "Empirical attempts at altering the χ^2 value by different slopes and intercepts of the line resulted in no further increase" (von Essen, 1960). (The "χ^2 value" will be a maximum if the off-diagonal cells, b and c, of the four-fold table are empty, i.e., the two partitions are completely consistent.)

This is yet another example of the errors that can occur when the investigator does not fully grasp the meaning of the statistical procedure he has deployed. It is quite similar to an earlier example in which it was demonstrated that the "survival transformation" of Poisson cell-survival data is completely equivalent to arbitrarily assigning a large weight to the observed response, m_1, at D=0, or, to imposing the non-stochastic constraint, $e^{\beta}0 \equiv m_1$. See Fig. 13a and Annex IV, part 6.

Such an investigation as von Essen's seeks a partition of the data such that the cases within a given category, say A, of the response attribute resemble each other (in respect of the values taken by the treatment variables) more than they do cases in the complementary category, say $\bar{A}$. The "line" that defines the optimal partition is known as the discriminant curve. We shall return to this issue below. But more important than this conceptual error in construction of a cross-classification represented by the four-fold tables is the fact that the existential question posed by von Essen is not the significance of the association - Does it differ from zero? - but rather its degree - Does it differ sufficiently from zero to be useful? A remark from a well-known textbook is much to the point: "A common mistake is to use the value of χ^2 itself as the measure of association. Even though χ^2 is excellent as a measure of the significance of the association [or in Fisher's locution, "Whether the observed departures from independence are or are not of a magnitude ascribable to chance"] it is not at all useful as a measure of the degree of association. The reason is that χ^2 is a function both of the proportions in the various cells and of the total number of subjects studied. The degree of association present is really only a function of the cell proportions. The number of subjects studied plays a role in the chances of finding significance if association exists but should play no role in determining the extent of association" (Fleiss, 1973). Moreover, "A measure of the degree of association between characteristics A and B which is derived from χ^2 but is free of the influence of total sample size [n] is the phi coefficient." ... "The phi coefficient is interpretable as a correlation coefficient" (Fleiss, 1973).

Thus, von Essen's analysis is wholly inconsistent with the inherent structure of the data to which it is applied and his inference - rejection of the null hypothesis of independence of the two attributes on the basis of the chi-squared statistic - is quite irrelevant to the issue of whether the subsequent deployments of the power-law models of the effects of tissue volume (area) on tissue dose-responses or, more precisely, on tissue isoeffects, produces useful, or even defensible, clinical results. But an answer to the latter question should be the "object of the exercise".

For the binary response variable, the degree, or strength, of association, ϕ, of the observed and estimated responses represented in a four-fold table can be estimated from the entries in that table: $\phi = (ad-bc)/[(a+b)(c+d)(a+c)(b+d)]^{1/2}$. Note that $\phi = \sqrt{}(\chi^2/n)$ and that ϕ is just the Pearson product-moment correlation, say R, of observed and expected responses. Note also that there are other measures of the degrees of association in four-fold tables such as the odds ratio, say OR, but for our immediate purposes, ϕ, or R, is preferable, although we have included estimates of the odds ratio and its standard error (Fleiss, 1973). Of course, $-1.0 \leq \phi \leq 1.0$ and values of ϕ, or R, close to zero imply very little association of the two dichotomies, whereas values close to one imply nearly perfect predictability of treatment response from a knowledge of treatment level. But, "... as a rule of thumb, any value [of ϕ] less than 0.30 or 0.35 may be taken to indicate no more than trivial association" (Fleiss, 1973). Note that, "The phi coefficient has a number of serious deficiencies, however; ... Carroll (1961) has shown that if either or both characteristics are dichotomized by cutting a continuous distribution into two parts, then the value of ϕ depends strongly on where the cutting point is set" (Fleiss, 1973). But, of course, the von Essen analyses are wholly dependent upon "... cutting a continuous distribution into two parts ..."

Although, as remarked above, von Essen did not summarize the major part of his evidence in the appropriate form of 2x2 tables we have constructed a set of such tables from his paper and present them at Tables 3, 4, 5 and 6. In a test of significance to reject the null hypothesis of independence with a probability of Type I error $\alpha = 0.05$, it is necessary that $\chi^2(1) > 3.84$. The respective chi-square statistics (Yates' correction) are 23.9, 6.32, 16.02 and 0.036. In Tables 3, 4, 5, and 6 the values of $\chi^2(1)$ reported by von Essen appear in parentheses following our estimate; e.g., in Table 3 we find $\chi^2(1) = 23.9$ (vs 21.0). We remark that if the test of significance were the appropriate inference on the data in the dichotomies then 1) the data in Table 6, Large Tumors, do not reject the null hypothesis and 2) a Fisher's Exact Test of significance is more appropriate than the chi-squared test for the data of Tables 4 and 6 owing to the small number (c=1) of observations in one of the cells.

But a test of significance is not relevant to the question at issue which is not the existence, but the degree of (linear) association between the two dichotomies; not χ^2, but ϕ is the appropriate statistic. For Tables 3, 4, 5 and 6 the values of ϕ are 0.234, 0.374, 0.202, and 0.092, respectively. Thus, on the criterion of the ϕ coefficient, the degree of association between level of treatment and level (binary) of response is trivial in these data (vide supra). Tables 3, 4, 5, and 6 give two measures of association: The ϕ-coefficient and the odds ratio (OR and OR$^{\#}$). Many authors prefer the latter as the most useful measure of association in 2*2 tables (Fleiss, 1973). However, for the above tables the odds ratio tells a story identical to that of the ϕ-coefficient. Only for Small Fields (Table 3) does the estimated odds ratio exceed its standard error by a factor of 2 or more. For Tables 4 and 6 the odds ratio is approximately equal to its standard error.

However, the distribution of the log odds ratio is more nearly Normal than is the distribution of the odds ratio itself (Kahn and Sempos 1989). Therefore, a better measure of precision of the estimate of association is given by the 0.95 CL on OR calculated as $\exp\{\ln OR \pm 1.96 se(\ln OR)\}$ where $se(\ln OR) = (a^{-1} + b^{-1} + c^{-1} + d^{-1})^{1/2}$. Thus, we have for Table 3: OR = 7.339 (3.119, 17.268), for Table 4: OR = 16.667 (1.705, 162.968), for Table 5: OR = 9.773 (2.763, 34.570) and for Table 6: OR = 2.121 (0.227, 19.848), where the 0.95 CL are in parentheses.

The log odds measure has another useful feature: For the logistic model, $\log\{\pi/(1-\pi)\} = \beta_0 + \beta_1 x$, where x is a dichotomous risk variable (x = 0, 1), we find that the maximum likelihood estimates, $\hat{\beta}_1$ and $Var(\hat{\beta}_1)$, are related to the odds ratio of the 2*2 table as $OR = \exp(\hat{\beta}_1)$ and $Var(\ln OR) = Var(\hat{\beta}_1)$. See Hosmer and Lemeshow, 1989.

Since the values of ϕ for three of the four tables do not even approach 0.30, and only one of the values of OR exceeds its standard error by a factor of 2 or more, we must conclude that both the slope and intercept of the respective discriminant curves are only poorly determined by these data. We note that this conclusion is supported by the statement of von Essen for the data on Necrosis/No Necrosis in Small Fields: "A number of slopes and intercepts appeared to give values of similar significance. The equations ranged from $D = 2500T^{0.20}$ to $D = 1400T^{0.31}$. The equation $D = 1950T^{0.27}$ appeared to give the most significant separation ..." But this slope (0.27) and intercept (1950) are just the simple average of the extremes. It seems likely that the intercept and slope of the discriminant curve (isoeffect lines) for both sizes of tumors are less well defined than they are for Small Fields and that even for Large Fields these estimates are not sufficiently

Table $\underline{3}$. Small Fields (< 10 cm^2). Necrosis ($\bar{N}$/N). Isoeffect Line: $\underline{D = 1950T^{0.27\#}}$.

		Predicted			
		$x \le x_0$	$x > x_0$		
Observed	$\bar{N}$	425	49	474	
	N	13	11	24	
		438		60	498

$\chi^2(1) = 23.91$(vs 21.0)

$p < 0.001$ (2-sided asymptotic)

$\phi = \sqrt{}(\chi_u^2/n) = 0.234$

OR = 7.339. se(OR) = 3.204. OR$^\#$ = 7.332. se(OR$^\#$) = 3.134.

von Essen estimate of level, π, of isoeffect line: $\pi^* = 13/438 = \underline{0.03}$.

$^\#$ $\underline{1950}$ = (2500 + 1400)/2; $\underline{0.27}$ = (0.20 + 0.34)/2 (?)

Table $\underline{4}$. Large Fields (> 10 cm^2). Necrosis ($\bar{N}$/N). Isoeffect Line: $\underline{D = 1700T^{0.27}}$.

		Predicted		
		$x \le x_0$	$x > x_0$	
Observed	$\bar{N}$	50	12	62
	N	1$^\#$	4	5
		51	16	67

$\chi^2(1) = 6.32$(vs 6.2). p = 0.002 (2-sided asymptotic)

$\phi = \sqrt{}(\chi_u^2/n) = 0.374$.

OR = 16.667. se(OR) = 19.389. OR$^\#$ = 12.120. se(OR$^\#$) = 12.047.

von Essen estimate of level, π, of isoeffect line: $\pi^* = 1/51 = \underline{0.02}$.

$^\#$ Requires Fisher's Exact Test: Hypergeometric probability = 0.010.

Table $\underline{5}$. Small Tumors (< 10 cm^2). Ablation ($\bar{A}$/A). Isoeffect Line: $\underline{D = 2680T^{0.14}}$.

		Predicted			
		$x > x_0$	$x \le x_0$		
Observed	A	289	138	427	
	$\bar{A}$	3	14	17	n_2
		292	152	444	

$\chi^2(1) = 16.02$ (vs 14.2). p = 0.0001 (2-sided asymptotic).

$\phi = \sqrt{}(\chi_u^2/n) = 0.202$

OR = 9.773. se(OR) = 6.30. OR$^\#$ = 8.660. se(OR$^\#$) = 5.231.

von Essen etimates of level, π, of isoeffect line: $\pi^* = 3/289 = \underline{0.01}$.

Table $\underline{6}$. Large Tumors (> 10 cm^2). Ablation ($\bar{A}$/A). Isoeffect Line: $\underline{D = 2950^{0.14}}$.

		Predicted		
		$x > x_0$	$x < x_0$	
Observed	A	14	33	47
	$\bar{A}$	1$^\#$	5	6
		15	38	53

$\chi^2(1) = 3.635{*}10^{-2}$ (vs 0.37). p = 0.849 (2-sided asymptotic)

$\phi = \sqrt{}(\chi_u^2/n) = 0.092$

OR = 2.121. se(OR) = 2.420. OR$^\#$ = 1.587. se(OR$^\#$) = 1.544.

von Essen estimate of level, π, of isoeffect line: $\pi^* = 1/14 = \underline{0.07}$

$^\#$ Requires Fisher's Exact Test: Hypergeometric probability = 0.328.

precise.

The 565 skin responses Necrosis/No Necrosis were coarsely stratified according to field size as Small Fields, $A_i < 10$ cm^2 (n=498) and Large Fields $A_i > 10$ cm^2 (n=67). The distribution-free discriminant curve estimates of the putative 0.03 isoeffect and ϕ coefficients are D2 = $1950T^{0.27}$ ($\phi = 0.24$) and D = $1700T^{0.27}$ ($\phi = 0.37$). Subsequently, a subset of size 12 of the Small Fields for which $A_i < 1$ cm^3, with No Necrosis at (D,T) = (3000, 1) were identified and stipulated in the von Essen 1960 paper to describe a third member of the family of 0.03 isoeffect curves for which the intercept was $D_3 = 3100$ rad.

The responses Ablation/Recurrence for 498 previously untreated skin <u>tumors</u> were coarsely stratified according to field size as Small Tumors, $A_i < 10$ cm^2 (n=444) and Large Tumors, $A_i > 10$ cm^2 (n=53). The distribution-free discriminant function estimates of the putative 0.99 isoeffect curves and ϕ coefficients are $D_1 = 2680T^{0.14}$ ($\phi = 0.20$) and $D_2 = 2950T^{0.14}$ ($\phi_2 = 0.10$). Subsequently, a subset of unspecified size of the large Tumors for which $A_i - 100$ cm^2 and for which the values of (D_i, T_i) were not disclosed were identified and stipulated to describe a third member of this family of $\pi = 0.99$ isoeffect curves for which the intercept was $D_3 = 3900$ rad. This is a rather apocryphal set of observations which were not otherwise discussed in the body of the von Essen 1960 paper and seem to appear in an ad hoc manner only in Fig. 10 of that paper (Figure 1 of this Appendix).

It should also be noted at once that the estimated <u>intercepts</u> of the isoeffect curves at A = 1 cm^2 (3100) and A = 100 cm^2 (3900) are not only the two largest but also are each quite different from those of the two other members of their respective families. For <u>skin</u> response we have $D_1 = 1700$, $D_2 = 1950$, $D_3 = 3100$ and $(D_2-D_1) = 250$, $(D_3-D_2) = 1050$. For <u>tumor</u> response we have $D_1 = 2680$, $D_2 = 2950$, $D_3 = 3900$ and $(D_2-D_1) = 270$, $(D_3-D_2) = 950$. In each case the difference between the first and second intercepts is less than 1/3 of that between the second and third. We shall discuss the consequences of this non-uniformity in the distribution of the estimated intercepts below.

It should also be noted that there is an evident asymmetry in the subsequent treatment of these two extreme "observations". Obviously, the two subsets of extreme observations at A < 1 cm^2 and A – 100 cm^2 include information on the radiation responses of <u>both</u> tumor and skin. However, in the construction of the isoeffect lines of Fig. 10 of the von Essen 1960 paper, this is ignored. For example, the range of <u>skin</u> field size is < 1 cm^2 < A < 30 cm^2. The isoeffect line for A = 100 cm^2 is omitted. Similarly, the range of <u>tumor</u> size is < 10 cm^2 < A < 100 cm^2. The isoeffect line for A < 1 cm^2 is omitted.

There are other idiosyncracies as well. For example how much information on skin response does the observations on a subset of size 12 (A < 1 cm^2 at a single level of treatment provide? For instance, if the conditional probability of Necrosis in the population, say θ, is high, say θ = 0.10, at (D, T) = (3000, 1) the probability of No Necrosis in a sample of size n = 12 is given by the last term of the binomial expansion: $(1-\theta)^n = 0.28$. If $\theta = 0.03$ as stipulated then $(1-\theta)^n = 0.69$. The observed response in the sample has a rather high likelihood for either level of θ in the population.

On examination of Fig. 10 in the von Essen 1960 paper it is found that the designation of the range of sizes represented in the data of each of the four double-dichotomies described in the body of the von Essen paper from which the estimates of the intercept and slope of two of the members of each family of isoeffect curves were obtained has been significantly altered. For <u>skin</u>, observations with $A_i < 10$ cm^2 are now further stratified into $A_i < 1$ cm^2 <u>and</u> 1 cm$^2 \le A_i$ < 10 cm^2; similarly, observations with $A_i > 10$ cm^2 are now divided into 10 cm$^2 < A_i \le 30$ cm^2. For <u>tumor</u>, the stratum $A_i < 10$ cm^2 is not further partitioned but the stratum $A_i > 10$ cm^2 is partitioned into observations with 10 cm$^2 \le A_i \le 30$ cm^2 <u>and</u> with $A_i - 100$ cm^2.

The discussion of Fig. 10 in the body of the von Essen (1960) paper also includes the odd, rather cryptic, disclaimer: "... only two of the isoeffect lines have been derived directly from the data. The construction of all other curves parallel to those two originally derived curves is, then, hypothetical." But which two curves does this disclaimer identify? Presumably, the only two

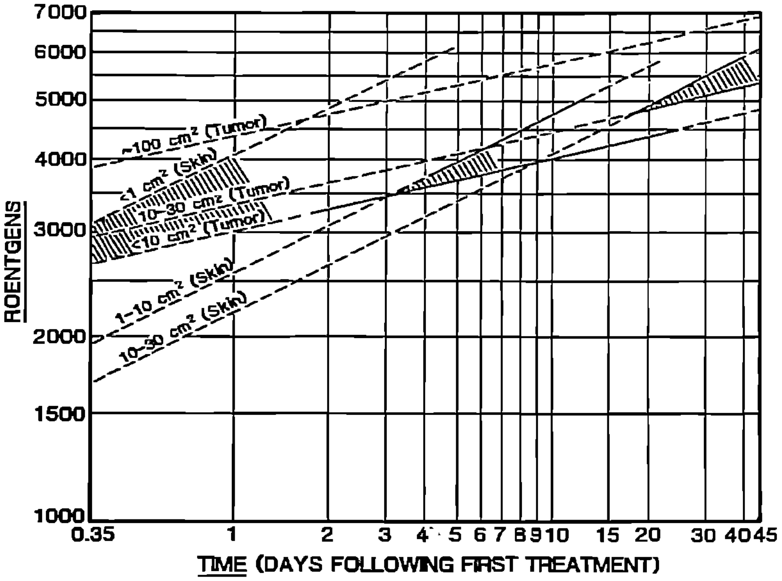


Fig. 1. "A hypothetical model based on the derived curves with assumptions that k varies while m and n remain constant in the equations $D = k_{skin}T^{n}$ and $D = k_{tumor}T^{m}$. The shaded areas refer to favorable therapeutic regions for particular treatment situations. For example, a lesion treatable by a field of S cm.2 may be treated with 3,450 r in 72 hours but may be more safely treated with a dose somewhat over 4,000 r in 240 hours."

"In addition, only two of the iso-effect lines have been derived directly from the data. The construction of all the other curves parallel to those two originally derived curves is, then, hypothetical." (von Essen, 1960).

The figure is reproduced, with permission, from Carl von Essen (1960).

isoeffect lines derived directly from the data are $D = 1950T^{0.27}$ (small fields, $A < 10$ cm^2) and $D = 1680T^{0.14}$ (small tumors, $A < 10$ cm^2) since these are the only curves for which the separation of the respective complementary classes is "highly significant" (See Tables 3 and 5). However, in neither case is the separation achieved "useful": $\phi = 0.23$ and $\phi = 0.20$, respectively. Moreover, these strata are still heterogeneous with respect to field and tumor size - the object of the study. Moreover, neither the curve for $A < 1$ cm^2 (Skin) nor the curve for $A - 100$ cm^2 (Tumor) were derived from large samples (since the data for $A < 1$ cm^2 (Skin) consists of the observed response in a subset of size 12 at only a single level of (D_i, T_i) and the data for $A - 100$ cm^2 (Tumor) are not disclosed at all).

It must be emphasized that the von Essen 1960 paper itself provides no estimates of the exponents of the volume (area) factors. It appears that the estimates of the exponents for the volume factors quoted by Cohen in his 1982 review (-0.27 and 0.14) are the result of some confusion of the respective exponents of the time factors, T, (0.27 and 0.14) which he ascribed to the volume (area) but which were not estimated in that paper. Misreading reports may occur rather frequently ("Misreading data seems to be common in studies ..." Hillcoat, 1984.)

The estimates of these exponents presented by von Essen in his 1963 paper appear to be obtained from a regression on the intercepts of the respective families of isoeffect curves defined by the stratification on area, A, for skin and for skin cancer. But we recall at once that in his 1960 paper von Essen remarks that the values of the intercepts for 2 out of 3 of the members of each family are "hypothetical".

In the von Essen 1963 paper the estimates of the levels of response for the isoeffect curves are $\pi > 0.99$ (tumor cure) and $\pi < 0.03$ (skin necrosis). The estimates of exponents of the time factors for the isoeffect curves are 0.12 (tumor cure) and 0.26 (skin necrosis). The estimates of the exponents for the volume (area) factors are given as 0.075 (tumor cure) and -0.16 (skin necrosis). The von Essen 1963 paper states that, "The slopes and intercepts were originally derived from data on the results of clinical radiation therapy of skin and lip cancer," and refers to the von Essen 1960 paper. No other descriptions of either the data or methods of analysis are provided. See Fig. 2.

It can be demonstrated that least squares estimates of the exponents of the volume (area) factors for tumor cure and skin necrosis based on the sets of intercepts of the two families of isoeffect curves in Fig. 10 of the von Essen 1960 paper are, respectively, -0.087, based on the areas $A_1 = 10$ cm^2 and $A_2 = 30$ cm^2, and -0.18 based on $A_1 = 1$ cm^2, $A_2 = 10$ cm^2, and $A_3 = 30$ cm^2. See Tables 7 and 8. The estimate of the exponent for the volume (area) factor for tumor cure is 0.16 based on the areas $A_1 = 10$ cm^2, $A_2 = 30$ cm^2, $A_3 = 100$ cm^2.

It thus appears that the von Essen 1973 estimates of the exponents were obtained from a different family of isoeffect curves than that from which the 1963 estimates were obtained. It can also be shown that these estimates are strongly influenced by the intercept at $A_1 = 1$cm^2 (skin necrosis) and $A_3 = 100$ cm^2 (tumor cure). Both intercepts appear to represent exceedingly dubious observations in the 1960 paper: A_1 represents a subgroup of 12 patients at a single level of the covariates (D,T) = (3000, 1). The data represented by A_3 (a 100 cm^2 tumor?) is not described at all in that paper.

There are several other aspects of the von Essen estimates and inferences which must be assessed more closely. The foregoing analyses were based upon the summaries of his data provided by von Essen. The following analysis must be based in part upon theoretical arguments and in part upon carefully chosen surrogate data since the primary von Essen data are not accessible. Nonetheless, because of the clinical importance of the matters at issue, these aspects must be addressed. The matter immediately at issue is the validity of the point estimates of the levels of response, π, described by the isoeffect curves, $D(\pi)$, of von Essen's papers and the Cohen (1982) review. However, larger issues of the proper construction and validation of dose response models of observational data - such as that of von Essen - must be considered as well.

von Essen's method of estimation of the location and slope of the isoeffect lines is described as follows:

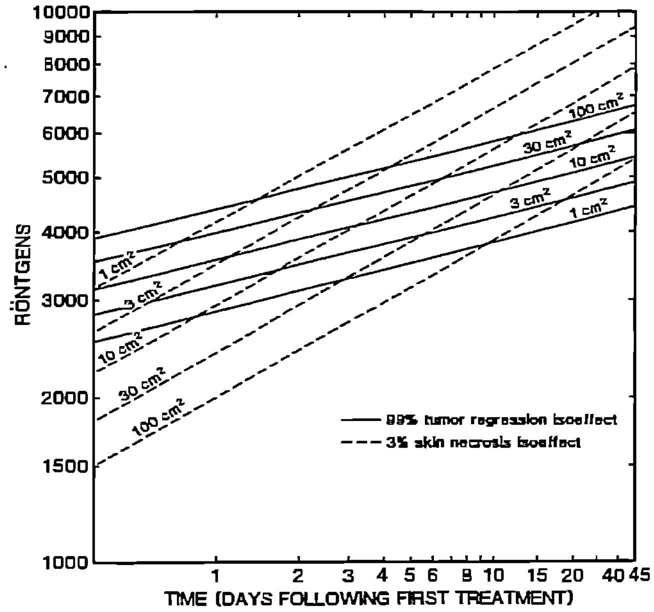


Fig. 2. "A graph with co-ordinates of cumulative dose in roentgens and elapsed time in days. The parallel iso-effect lines for skin tolerance and tumor cure vary with area." (von Essen, 1963).
The figure is reproduced, with permission, from Carl von Essen (1963).

Table 7. Regression Model of Volume Effect. $y = \beta_0 + \beta_1 x$.

Necrosis of Skin.

$y = \log_{10} D(\pi)$. $x = \log A_i$

$n = 3$: $A_1 = 1$ cm$^{2\#}$, $A_2 = 10$ cm^2, $A_3 = 30$ cm^2

$$\begin{array}{ll} y = 3.486 & - 0.180x.\ \bar{R}^2 = 0.979 \\ (182.707) & (-9.738)^{\#\#} \end{array}$$

Subgroup of 12 patients at T = 0.35, D = 3000 rad, $\pi = 0.00$ (No Necrosis).

$\hat{\beta}_j / \sqrt{\mathrm{Var}(\hat{\beta}_j)}$

Table 8. Regression Model of Volume Effect. $y = \gamma_0 + \gamma_1 x$.

Ablation of Tumor

$y = \log_{10} D(\pi)$. $x = \log A_i$

i) $n = 3$: $A_1 = 10$ cm^2, $A_2 = 30$ cm^2, $A_3 = 100$ cm$^{2\#}$

$$\begin{array}{ll} y = 3.252 & + 0.164x.\ R^2 = 0.879 \\ (50.52) & (3.943) \end{array}$$

The "100 cm^2 tumor" at T = 0.35, D = 3900 rad, $\pi = 1.00$ (Ablation) (? These data are not described in the main body of the 1960 paper.)

ii) $n = 2$: $A_1 = 10$ cm^2, $A_2 = 30$ cm^2

$y = 3.341 + 0.087x$. $R^2 = 1.000$

No other statistics

i) "A simplified derivation of isoeffect curves by empirical fitting of smooth curves relating a certain probability of cure to the coordinates of dose and time is attempted. The effectiveness of the curves to separate cured and non-cured lesions is tested by Chi-squared analysis."
ii) "Empirical attempts at altering the χ^2 value by different slopes are intercepts of the line resulted in no further increase."
iii) "No empirical alternations of slope and intercept improved the value" [of the chi-squared statistic]. Referring to Table 1 it is evident that χ^2 is maximized by adjusting the slope and intercept of "the line" to reduce the sizes, b and c, of the off-diagonal cells. But this is equivalent to reducing the misclassification rates achieved by the linear discriminant function implied by the maneuver. (See Cornfield, 1962; Truett et al, 1967; Halperin et al, 1971; Harris, 1975; Press and Wilson, 1978.) That is, although von Essen does not identify or describe any further his method of estimation of the slopes and intercepts of his isoeffect curves, it seems clear from the foregoing remarks that his isoeffect curves are, in fact, the linear discriminant curves that optimally separate the two sets of responses in each case: ablation and recurrence (tumors); necrosis and non-necrosis (skin). However, a discriminant curve corresponds to the π = 0.50 isoeffect curve, not to the π = 0.99 isoeffect curve (tumor ablation) and not to the π = 0.03 isoeffect curve (skin necrosis), as we shall now demonstrate.

The measure of the level of tumor and tissue responses - the effect - that are described by the respective isoeffect lines is incorrectly specified in the von Essen papers as the ratio of the fractions of the two populations "above the line", or "below the line", respectively. To see this more clearly, let us assume that the density functions of the conditional joint distributions of dose and days, say $\underline{x}^T$, on the two complementary groups, say $\bar{E}$ and E, are $f_1(\underline{x}^T) = f_1$ and $f_2(\underline{x}^T) = f_2$, respectively. Then the conditional probability of occurrence of the effect, say E, at a given level of dose and time is $P(E|\underline{x}^T) = P(E \text{ and } \underline{x}^T)/P(\underline{x}^T)$ or $P = n_2f_2/(n_1f_1 + n_2f_2) = [1 + (n_1/n_2)(f_1/f_2)]^{-1}$. However, it can be shown that the position of the line that gives the best separation of the members of the two complementary groups, $\bar{E}$ and E, is defined by the relation $n_1f_1 = n_2f_2$. Obviously, along this line the probability of response is P = 0.50. Note that although the two conditional bivariate distributions of D and T may not be bivariate Normal this is, perhaps surprisingly, often a fairly good approximation (See, for instance Herbert, 1986a and 1986d).

Thus, it seems safe to assume that the isoeffect curves of von Essen define quantiles, $D(\pi)$, of the respective responses that are each much closer to π = 0.50 than to either π = 0.03 (skin necrosis) or to π = 0.99 (tumor ablation) that have been described in the von Essen papers and widely quoted since. It can be seen from the Tables 3-6 that the respective 0.50 isoeffect curves have the anticipated dependence on volume (or area) for each end-point, ablation or necrosis.

As remarked above, since von Essen did not publish his data - and it is not otherwise assessable - it is impossible to demonstrate these effects for his model. However, a surrogate data set will serve to illustrate the principles at issue. Figure 3a shows a scattergram in the dose-time plane of the responses for Hodgkin's disease treated by irradiation. There are 14 recurrences and 32 ablations. The total sample size (n=46) is comparable to von Essen's samples for Large Fields (n=67. Table 4) and Large Tumors (n=53. Table 6). Note that if the von Essen data for Small Fields and Tumors were further stratified according to field size, the sample sizes of the more homogeneous data would no doubt be much smaller than those of Tables 3 and 5. (Unless the data are homogeneous with respect to volume, the effects of the covariates upon response are confounded.) The respective 0.80 contour ellipses are super-imposed. These ellipses are constructed on the assumption that the joint distribution of dose and time is bivariate Normal on each response. Since the respective ellipses include 24/32 = 0.75 of the ablations and 11/14 = 0.79 of the recurrences, this suggests that each of the respective distributions is consistent with the assumption of bivariate Normality. This evidence is corroborated by the respective chi-squared probability plots in Fig. 4.

Parallel analyses of these data according to two different dose-response models are presented in Table 9a. Table 9ai presents the bivariate linear discriminant model of these data. This is the dose-response model implied by the method of estimation of isoeffect curves in the von Essen 1960

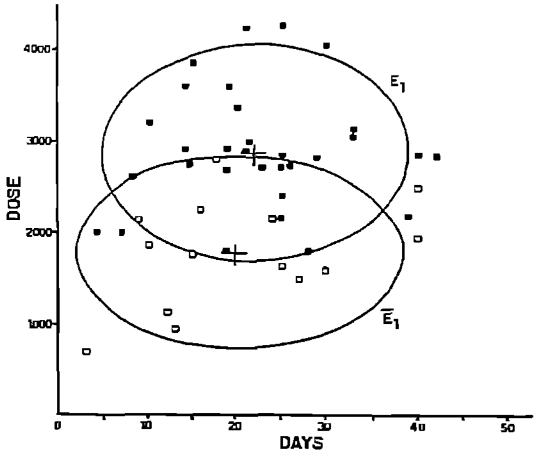


Fig. 3a. Super-position of the respective 0.80 contour ellipses for the conditional distributions of D,T on the response E_1 (ablation) and $\bar{E}_1$ (recurrence). The respective numbers are $n(E_1) = 32$, $n(\bar{E}_1) = 14$. The closed symbols identify ablations; the open symbols identify recurrences. The data are a 10% stratified random sample of the Hodgkin's disease data of Friedman and Pearlman, 1967. If the conditional distributions are bivariate Normal there should be $0.20^*n(E_1) - 6$ and $0.2^*n(\bar{E}_1) - 3$ observations outside the respective ellipses. The figure discloses that there are 8 and 3 such observations respectively, suggesting that the conditional distributions are indeed bivariate Normal.

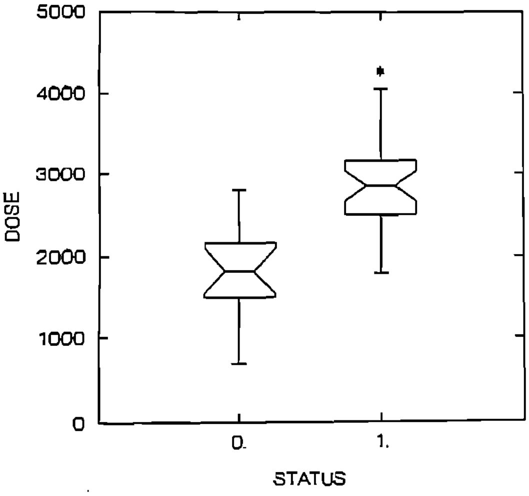


Fig. 3b. Box plots of the respective conditional distributions of dose on status: ablation = 1, recurrence = 0. In the Box plot the horizontal line at the narrowest part of the central polygon represents the median (the 0.50 quantile) of the distribution. The lower and upper limits of the polygon represent the 0.25 and 0.75 quantiles, respectively. The upper and lower limits of the notch represent the 0.95 confidence limits on the median. The vertical lines include all of those observations within 1.5* interquartile range from the median. Observations that lie between 1.5 and 3 times the interquartile range from the median are identified by an asterisk; observations that lie beyond 3 times the interquartile range are identified by a circle. It appears that the respective dispersions are equal and that the overlap of the two distributions is not so large that some discrimination between the two responses on the basis of dose is not possible; that is, the estimated coefficient of the dose in the cognate dose-response model will exceed its standard error. A characteristic weakness of clinical binary response data is that the degree of overlap of the two conditional distributions of treatment variables is so large that the two conditions cannot be discriminated on the information in the treatment variables and hence the parameter estimates of the dose-response model do not differ significantly from zero.

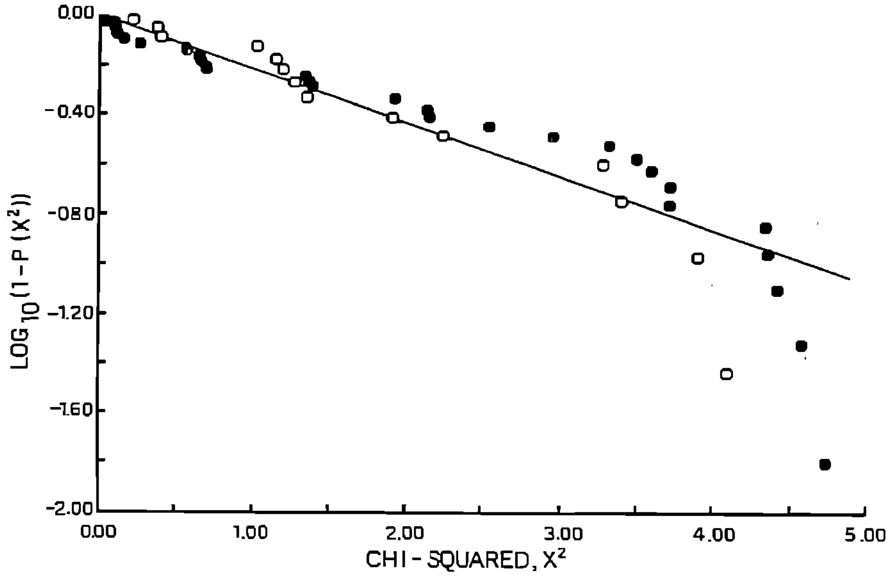


Fig. 4. A graphical chi-squared test of bivariate Normality (Hald, 1952) of the conditional distributions of Fig. 3a. If the conditional distributions are random samples from a bivariate Normal distribution the respective sets of points will lie close to the theoretical straight line - as seems to be the case with these data. If the two conditional distributions are bivariate Normal, with equal dispersions, the parameters of a dose-response model can be estimated by the coefficients of the corresponding discriminant function on the basis of Bayes' Theorem.

Table 9a. Parallel Analyses of Surrogate Data (Hodgkin's Disease) n=46. k=3.

i) Discriminant Function Model

$P(E_1|\underline{x}^T) = [1 + e^{-z}]^{-1}$

$z = \alpha_0 + \alpha_1 D + \alpha_2 T$

$z = - 5.055 + 2.684{*}10^{-3}D - 0.0174$

$\qquad\qquad (5.292) \qquad\qquad (0.522)^{\#}$

$D_m^2 = 2.908.\ R = 0.626$

$\hat{\alpha}_j/\sqrt{Var}(\hat{\alpha}_j).\ j = 1,2$

ii) Probit Regression Model

$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$z = \phi^{-1}(\pi),\ x_1 = \log D,\ x_2 = \log T$

$z = -32.130 + 10.049x_1 - 0.860x_2$

$\quad (-3.390) \qquad (3.486) \quad (-0.860)\#$

$\chi^2 = 33.767.\ P(\chi^2 > \chi_c^2|43) = 0.842.$ Do not reject H_0.

$\hat{\beta}_j/\sqrt{Var}(\hat{\beta}_j).\ 0 \le j \le 2.$

Table 9b. Parallel Analyses of Surrogate Data (Hodgkin's Disease). n=46. k=2.

i) Discriminant Function Model

$P(E_1|D) = [1 + e^{-z}]^{-1}$

$z = \alpha_0 + \alpha_1 D$ (rad)

$z = -5.306 + 2.628{*}10^{-3}D.\ D(0.50) = 5.306/2.630{*}10^{-3} = 2018$ cGy (See Fig. 8).

$\qquad\qquad (5.296)^{\#}$

Mahalanobis distance, $D_m^2 = 2.880.\ R = 0.620$

$\hat{\alpha}_j/\sqrt{Var}(\hat{\alpha}_j).\ j=1$

ii) Probit Model

$z = \beta_0 + \beta_1 x_1$

$z = \phi^{-1}(\pi),\ x_1 = \log D$ (rad)

$z = -31.737 + 9.604x_1.\ D(0.50) = 10^{31.737/9.604} = 2016$ cGy.

$\quad (-3.352) \quad (3.401)^{\#}$

$\chi_c^2 = 33.359.\ P(\chi^2 > \chi_c^2|44) = 0.879.$ Do not reject H_0.

$\hat{\beta}_j/\sqrt{Var}(\hat{\beta}_j).\ j=1,2$

paper. Note, however, that the graphical tests of the bivariate Normality of the conditional distributions of the covariates in Figs. 3 and 4 require that the model be constructed on the untransformed covariates: $\underline{x}^T = (D,T)$.

The Mahalanobis distance, $D_m^2 = 2.88$, provides an assessment of the degree of separation of the two complementary groups achieved by the use of the discriminant function (Good discrimination is achieved only for $D_m^2 > 6.0$). R is the multiple correlation coefficient for the binary regression model of the discriminant function (R is a function of D_m^2). The analysis disclosed that the coefficient of T is not significant: $\hat{\alpha}_2/\sqrt{Var(\hat{\alpha}_2)} < 2.0$ and hence the covariate T should be deleted from the model as it has been. See Fig. 5. Note however, that the discriminant function model yields only asymptotic estimates of $Var(\hat{\beta}_j)$, $j = 1,2$. For small samples such estimates are dubious.

Table 9aii presents the probit model of these surrogate data. Note that in this model the logarithms of dose and time appear rather than the raw values. However, a probit model was also constructed on the raw values of dose and time and it was found that there was no difference in either the goodness-of-fit, measured by the Pearson chi-squared statistic, χ_c^2, or in the precision of estimate, measured by the ratios, $\hat{\beta}_j/\sqrt{Var(\hat{\beta}_j)}$, $0 \leq j \leq 2$ from the probit model constructed on the logarithms. However, there is some reason to believe, a priori, that in tissue response to radiation, the covariates dose and time combine multiplicatively rather than additively, and hence that the probit model based on their logarithms is preferred. Note that in the probit model also the coefficient of T is not significant: $\hat{\beta}_2/\sqrt{Var(\hat{\beta}_2)} < 2.0$. The value of the Pearson chi-squared statistic suggests that the model is not rejected by the data although the value of $\hat{\beta}_2/\sqrt{Var(\hat{\beta}_2)}$ signals that the model may be over-fit and hence that the time factor should be deleted.

Note that the use of the Pearson chi-squared test of concordance is questionable when the numbers at risk, n_i, at $\underline{x}_i^T$, become small, as in the present case, in which $n_i=1$, $1 \leq i \leq 46$. However, so long as $\chi_c^2 \leq$ df, the sample provides no evidence for rejecting the model of such data. For such ungrouped data the parallel linear discriminant model should be constructed to check the overall consistency of such a model with the data since the graphical tests of the assumptions, presented in Figs. 3 and 4, for this model are, of course, valid for such data; it is just such data that comprise the discrimination problem.

Table 9b presents the parallel analyses of the surrogate data for the two dose-response models from which the time factor has been deleted. The respective dose-response curves are shown in Figs. 5 and 6. Comparison of these figures discloses what may be termed an inherent difference in the respective statistical infrastructures of these two models: For the linear discriminant model the response is estimated by a function of the ratio of the ordinates of the respective density functions of the conditional distribution of observed dose, the "discriminating variate", on the two complementary responses (ablation/recurrence) whereas for the probit model the response is estimated by the ordinate of the distribution function of the unobservable tolerance doses (the "area under the curve"). For a single discriminating variable, D, the dose-response model obtained from the two-group discriminant function is $\log[\pi/(1-\pi)] = \beta_0 + \beta_1 D$ in which the parameter estimates are simple functions of the aggregate statistics $(\bar{x}_j, s_j^2)$, $j = 1,2$, of the two conditional distributions:

$$\hat{\beta}_1 = (\bar{x}_2 - \bar{x}_1)/s^2 = 2.628*10^{-2}$$

$$\hat{\beta}_0 = -\log[(1-\pi)/\pi] - \hat{\beta}_1(\bar{x}_1 + \bar{x}_2)/2 = -5.302$$

where $s^2 = [(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1 + n_2 - 2)$ on the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$. The asymptotic estimate of the standard error of the slope is $\sqrt{Var(\hat{\beta}_1)} = \sigma^{-1}(n_1^{-1} + n_2^{-1}) = 4.962*10^{-4}$. See Halperin et al, 1978.

Figure 7 presents a super-position of the respective dose-response curves of the discriminant function and probit models of Table 9b. Note that for $\pi > 0.30$ the two curves are nearly coincident, despite the fact that the structures of the two models are quite disparate.

Figure 8 presents a scattergram of the dose-response data on Hodgkin's disease on which the discriminant curve, that is, the $\pi = 0.50$ isoeffect curve, is super-imposed. In the estimation procedure that was used in the von Essen (1960) paper the level, π, of response (effect) is

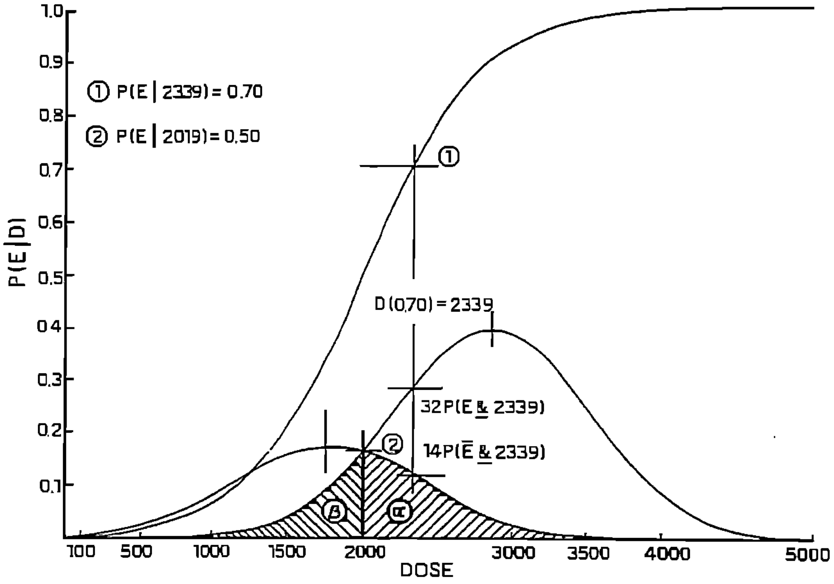


Fig. 5. The diagram describes the construction of the dose-response curve for the logistic model of the data of Fig. 3a by Bayes' Theorem:

$$P(E_1 | D) = \left[1 + \frac{n(\bar{E}_1)P(\bar{E}_1 \text{ and } D)}{n(E_1)P(E_1 \text{ and } D)}\right]^{-1}$$

Note that the numerator and denominator of the ratio are just the ordinates of the respective density functions of the two conditional distributions of dose, D. The estimates of the parameter vector of the model are given by the coefficients of the discriminant function for the two distributions. See Table 9a. Obviously, $P(E_1|D) = 0.50$ at that level of D for which $n(\bar{E}_1)P(\bar{E}_1$ and D) = $n(E_1)P(E_1$ and D); this is the point at which the two density functions intersect. For these data $P(E_1|2019) = 0.50$. The areas α and β are measures of false positive and false negative error rates, respectively. It can be shown that α + β is a minimum at this level of D which defines the discriminant curve for the discriminant function model.

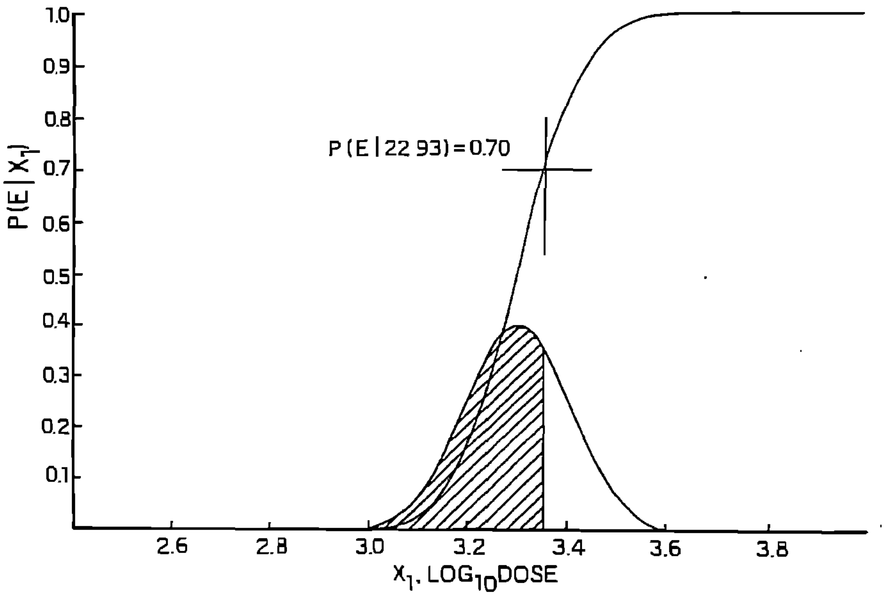


Fig. 6. The dose-response curve for the probit model of the data of Fig. 3a on the log dose metameter. The density function is that for the tolerance distribution of log dose. See Table 9b. Note that the estimate of the level of response is described by the area under the curve of the density function of the tolerance distribution in the probit model, but by the ratio of the ordinates of the respective density functions of the two conditional distributions in the logit model.

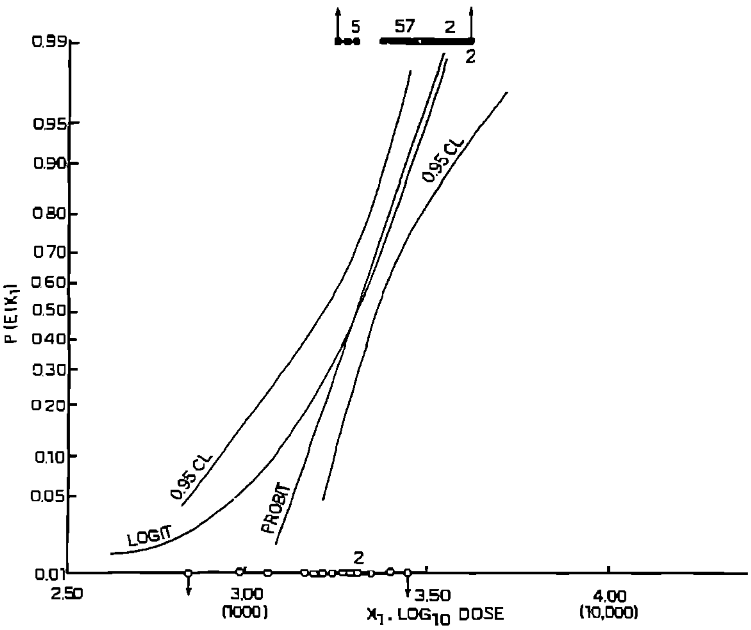


Fig. 7. Super-position of the dose-response curves obtained from the logit and probit models of the data of Fig. 3a. It is obvious that the two models are equivalent in the region of observed levels of response even though the dose metameters are quite different: D(logit) vs logD(probit).

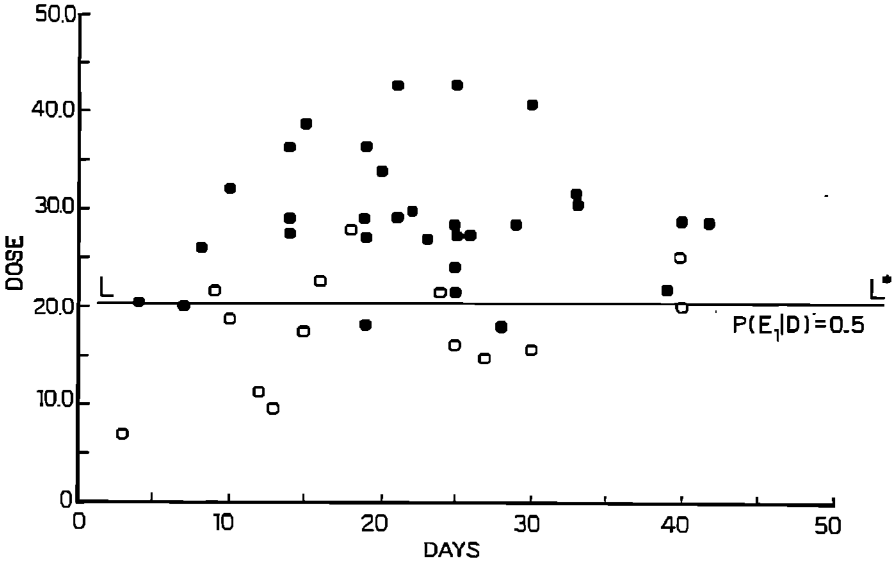


Fig. 8. The linear discriminant curve L-L* for the data of Fig. 3a. L-L* corresponds to point 2 in Fig. 5. von Essen's cross-classifications of irradiated tumors and tissues into numbers of responders/numbers of non-responders and numbers above the line/ numbers below the line (see Tables 3-5) were apparently motivated by a discriminant analysis criterion similar to that shown in Fig. 5; that is, he adjusted "by eye" the position and orientation of "the line" to minimize the number of misclassifications - as is done analytically in a formal discriminant analysis. But, as we have shown, the discriminant curve coincides with the 0.50 isoeffect curve, not the 0.03 or 0.99 isoeffect curves reported by von Essen.

estimated by the ratio of 5 (of 14) recurrences and 28 (of 32) "cured lesions" that are above the line. Therefore the recurrence rate is $\pi = 5/28 = 0.179$ or 18 percent recurrence. The von Essen procedure would incorrectly identify this curve as the 0.82 isoeffect line.

Table 10 presents the comparison of the observed statistics with that estimated from the model. For the four-fold table $\chi^2 = 10.892$, ($p = 0.001$) and Pearson's phi coefficient is $\phi = 0.620$. (The respective misclassification rates are $\alpha = 5/14 = 0.357$; $\beta = 5/32 = 0.156$.)

Although we have shown that the model of dose-response constructed from the discriminant function gives estimates of response consistent with those obtained from the more familiar probit regression model (See Fig. 7) it will be useful to demonstrate that the discriminant function model of dose-response gives similar results to those of a logistic regression model of dose-response. If we construct a logistic regression model, $\log[\pi/(1-\pi)] = z = \beta_0 + \beta_1 D$, of the surrogate Hodgkins data on the dis-aggregated data with dose, D(rad), we obtained the estimated logit, $z = -6.519 + 3.189{*}10^{-3}D$, with respective t-statistics of -2.983 and 3.257. Again, the 0.50 isoeffect curve is given by $D(0.50) = 6.519/3.189{*}10^{-3} = 2044$ cGy. The parameter estimates and standard errors are obtained by maximum likelihood methods. The Hosmer-Lemeshow chi-squared statistic is 5.339 on 8 df ($p = 0.72$) and McFadden's $\rho^2 = 0.422$, indicating a very good fit of the model and data. From these parameter estimates, $(\hat{\beta}_0, \hat{\beta}_1)$, we now construct a dichotomous "dose" variable, x, with $x = 1$ "above the line" and $x = 0$ "below the line", where the "line" is just the 0.50 isoeffect curve, $D(0.50) = -\ \hat{\beta}_0/\hat{\beta}_1$, estimated from the model. Then a logistic regression model on the dichotomous variable, x, $z = \beta_0 + \beta_1 x$, gives an estimate of the odds ratio, $OR = e^{\hat{\beta}}1 = 9.720$, with 0.95 CL of (2.278, 41.481). But these estimates, constructed from the dis-aggregated data, are identical to those obtained from the discriminant function model, (See Table 9bi) in which the model parameters are estimated from the aggregate statistics (means and variances) of the two conditional distributions. Note also that the point and interval estimates of the odds ratio for Table 10 are quite similar to those to Table 5 for the von Essen data. We think that this comparison strengthens our conjecture that the von Essen "0.03" and "0.99" isoeffect curves are, in truth, the respective discriminant curves ("0.50 isoeffect curves") for the four sets of observations: small fields, large fields, small tumors, large tumors.

The line which best separates the two groups of responders, the discriminant curve, is the 0.50 isoeffect line. It is $\hat{D}(0.50) = 2017$ rad. The 0.99 isoeffect line is at 3764 rad. This analysis of the surrogate data strongly suggests that the levels of response, π, presented in the von Essen 1960 paper are biased. Therefore, the levels of response, π, that might be achieved in practice may differ by consequential amounts from those anticipated on the basis of the model. In the present case the two estimates of D(0.99) differ by almost a factor of two (3764/2017 = 1.87). Moreover, it must be recalled as well that the confidence limits on the quantiles, $D(\pi)$, of a dose-response curve - or surface - increase dramatically (quadratically) as π increases (or decreases) from $\pi = 0.50$ so that the confidence limits on extreme quantiles such as for $\pi = 0.03$ and $\pi = 0.99$ are exceedingly broad. This is suggested in Fig. 9 for the probit model of the surrogate data. Even for the response level $\pi = 0.95$, the confidence limits extremely wide: $D(0.95) = 2988$, $P(2590 \leq D(0.95) \leq 4471) = 0.95$. That is, the isoeffect dose may lie within a range of about 0.80 to 1.50 of that estimated from the model.

Given the evidence of these secondary analyses, one must ask whether the empirical findings on "volume effect" that were reported in the 1960 and 1963 papers by von Essen and were subsequently interpreted in the 1982 review by Cohen and cited in the 1988 review by Withers, et al as evidence for the existence of a volume effect on the radiation responses of irradiated tissues believable? It would seem that they are not:

1) The findings were presented in the published reports as a test of significance (rejection of the null hypothesis of no association) but this is the right answer to the wrong question (a Type III error); the matter at issue required a measure of the strength, not the existence, of an association. The secondary analysis disclosed that a) the use of the χ^2 statistics is incorrect and b) that the actual associations in question based on the appropriate related measure, $\phi = \sqrt{\chi^2/n}$, were of a trivial degree.

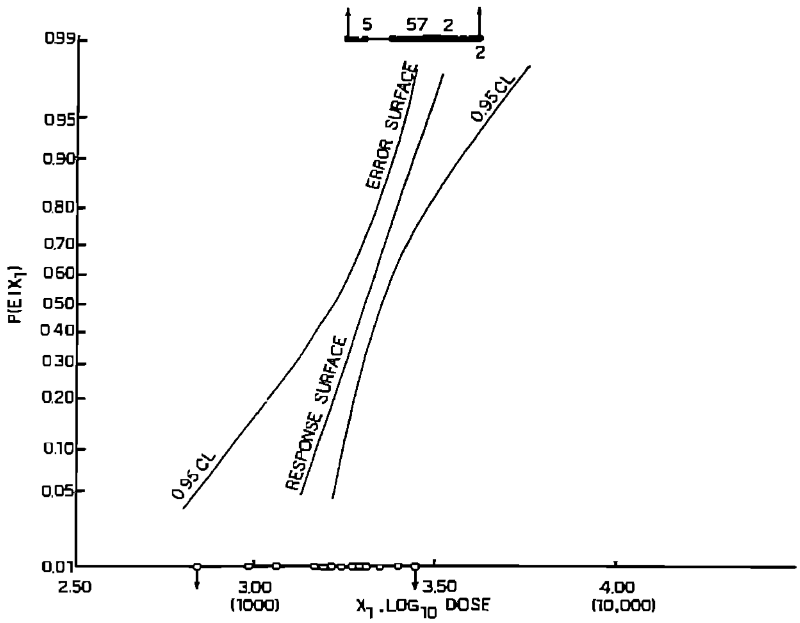


Fig. 9. Probit dose-response curve and 0.95 confidence limits (fiducial limits, FL) for the data of Fig. 3a.

Table 10. Surrogate Data. Hodgkin's Disease. Abaltion $(\bar{E}_1/E_1)$. Discriminant Function Model.

		Predicted		
		D > D(0.50)	D ≤ D(0.50)	
Observed	E_1	27	5	32*
	$\bar{E}_1$	5	9	14
		32	14	46

$\chi^2(1) = 10.882$. $p = 0.001$

$\phi = \surd(\chi_u{}^2/n) = 0.487$ (=R)

OR = 9.720. se(OR) = 7.193. OR$^\#$ = 8.636. se(OR$^\#$) = 6.140.

"von Essen estimate" of level, π, of isoeffect line: $\pi^* = 5/28 = 0.18$. $1 - 0.18 = \underline{0.82}$ "isoeffect line".

* In Figure 8 it appears that there are only 4 (rather than 5) ablations below the D(0.50) isoeffect line since there are two observations at (D, T) = (19.97, 7).

2) The levels of binary tissue response that were reported in the published papers were $\pi = 0.03$ (skin necrosis) and $\pi = 0.99$ (tumor ablation). But our secondary analyses disclosed that both of these levels were, instead, probably very close to $\pi = 0.50$.
3) The location of the intercepts of 2 (of 3) isoeffect curves for each tissue, skin and skin cancer, are, as von Essen acknowledges, "hypothetical". That is, most of the "data" from which the estimates of the exponents of the areas are subsequently estimated, seem to be fictitious - non-experiential - suggesting that data qua data were not taken too seriously in either this investigation or in those reviews that cited it. This is an attitude that seems to be not uncommon in the field. We have described other instances of it in section 12 and in the Annexes II-IV of this report.

We note that it is only because the findings of the 1960 paper of von Essen are summarized, quite inappropriately, as significant values of chi-squared statistics for 2x2 tables that the reported results might appear to be "believable" to the casual reader. This reporting practice appears to be an instance of a motivational error induced by the well-known and wide-spread "prejudice against the null hypothesis" which results in papers that report positive - "significant" - results being much more likely to be submitted and accepted for publication than those reporting negative - non-significant - results. This practice leads to the "publication bias" that has been found to severely degrade the clinical value of the literature on randomized controlled clinical trials. (See sections 2.2 and 5.1-5.3.)

In the present context it is, perhaps, of interest to recall that in the issue of 30 June 1988 the British journal Nature published a study that reported what were acknowledged to be, in the published remarks of the editor John Maddox, "unbelievable results." The publication of this paper, "has generated a good deal of publicity and raised a few eyebrows in the scientific community ..." presumably a good deal - but, to be sure, not all - of the commotion has arisen because it is widely held to be the common - and normative - practice of scientific journals to publish only results that are, on the evidence of peer-review, accepted as "believable", that is, are "capable of instilling faith, trust or acceptance (a believable explanation)" (Webster's Third New International Dictionary (Unabridged) 1967). The von Essen papers as well as several others listed in Table 1 suggest that faith in the peer-reviewed literature may not infrequently be misplaced. See also Williamson et al, 1986.

What Should We Do Now?

The existence of a volume effect is plausible: 1) There appears to be clinical evidence for it; 2) Several physico-chemical failure phenomena show the existence of a "size-effect" (Herbert 1978, 1983, 1985). That is, "It is commonly observed in the study of the strength of materials that the larger specimens will fracture under less applied stress, breakdown under less applied voltage, corrode in a shorter time, etc." (Herbert, 1985). The extreme value theory that has been developed to account for size-effects in other failure phenomena can also account qualitatively for the volume dependence of both normal and tumor tissue responses to ionizing radiation (Herbert 1982, 1985). Schultheiss, Orton and Peck (1982) have also presented a very cogent theoretical analysis of the volume effect in normal tissues. (It should be remarked that the 1983 and 1985 papers by Herbert, both based on the statistical theory of extreme values (See Bury, 1975) seem to be the first reports to introduce and then elaborate into a coherent model of "volume effects", the series and parallel connections of tissue elements that now enjoy considerable currency in such studies. See Withers et al, 1988a and Niemerko and Goitein, 1992. Both of these later papers present very interesting and elegant treatments of some aspects of this problem.)

The foregoing analyses suggest one or two ways to obtain less ambiguous empirical evidence for the existence of volume effects in the clinical responses of irradiated tissues to ionizing radiations. Both are based on the dose-response model rather than the isoeffect model. 1) Stratify the target tissues according to size (volume) and identify the m strata by (m-1) indicator variables, 2) construct a probit dose-response surface, say $z = \underline{x}^T\underline{\beta}$, including terms representing interactions of volume with treatment variables.

Another method is to simply include the volume, V_i, of tissue(s) irradiated at $\underline{x}_i^T = (x_0, x_1, x_2, x_3)$ or a function thereof, as a covariate, say $f(V_i)$, in a probit model, say $z_i = \beta_0 + \beta_1 x_{1i}$

$+ \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 f(V_i)$, $1 \leq i \leq z$, where $x_1 = \log D$, $x_2 = \log T$ and $x_3 = \log N$. If the ratio $\bar{\beta}_4/\sqrt{Var(\bar{\beta}_4)} > 2.0$, then the concept of a volume effect cannot be rejected - by these data.

Appendix II. A Somewhat Different Perspective on Radiation "Tolerance."

"The question is," said Alice, "whether you can make words mean so many different things." "The question is," said Humpty Dumpty, "which is to be master - that's all."

L. Carroll, Through the Looking-Glass

The debate on the issues of the nature of - and the possible role(s) for - the concept of "tolerance" in radiation therapy has always been somewhat theological and even divisive, a bit like those on the issues of Transubstantiation vs Consubstantiation - or on whether it really is "less filling" rather than simply "tastes great":

1) "The tissue tolerance dose is a minimal dose, in that most clinicians are only willing to accept a low lethal complication rate in normal vital tissues ... The suffering due to excessive therapy is rarely more tolerable than the suffering due to the disease."

P. Rubin and G. Cassaret, 1968

2) "The radiation therapist is admittedly treating to 'tolerance' doses rather than specific tumoricidal doses. The risks are poorly defined, and therapeutic ratios remain largely abstracts rather than concrete estimates. The amount of radiation damage that is acceptable for the purpose of curing a cancer remains one of personal philosophy as long as the overlapping zones of normal tissue tolerance and tumor curability are inadequately defined."

P. Rubin and G. Casarett, 1972

3) "Tolerance as a term characterizing a level of radiotherapeutic effect that has outlived its usefulness."

R. Andrews, 1986

4) "Tolerance is an important and valid concept in radiotherapy. It is relevant to every treatment decision. Yet remarkably, while the general thrust of the concept is well understood the exact definition remains obscure and couched in subjective terms such as "severe" and "acceptable".

G. Jones and E. Laukkanen, 1987

5) "Tolerance dose may be defined as that which produces an acceptable probability of a treatment complication."

Withers et al, 1988a

We maintain that the term and the concept of tolerance only need be more explicitly and concisely defined and implemented by a statistically adequate model in order to be understood and to retain whatever usefulness it might have in the context in which it usually arises. It is, moreover, a fructifying concept; as we shall show, it is productive of new perspectives on the design of experiments. In their 1987 letter responding to Andrews (1986) comments anent tolerance, Jones and Laukkanen state that, "The challenge for the future is to formulate a model [of tolerance] for acceptance." We describe a novel response to that challenge in the present exposition.

The "tolerance" implied in most discussions of the issue in radiation oncology is, as Rubin and Casarett have observed, that of the radiation oncologist: the site of the "radiotherapeutic effect" at issue lies in the cognitive centers of the attending radiation therapist. Withers, in his 1989 Failla Lecture, updates and elaborates this view: "There is no such thing as a tolerance dose for a tissue or organ, only for a tissue or organ in a given set of clinical circumstances. The tolerance dose had been defined as the regimen that produces the maximum acceptable probability of complications in a given treatment situation. What is tolerable for one patient may be intolerable for another, acceptability being subjective. In experimental radiobiology, the term is meaningless because the animal has no say in what is tolerable: informed consent does not involve the animal, its spouse or blood relatives, nor are legal consequences likely from misjudgment.

Tolerance dose in the clinical situation is a complex and flexible value. ... The virtue of the term 'tolerance' is as an abbreviation for a complex and patient-specified risk-benefit equation, in which usage it is extraordinarily valuable."

Thus, the observed response in the irradiated tissues of the patient must be viewed as merely a proxy for that of the physician. The prediction of the decision to continue treatment, or not, in any given case, can be described empirically by exactly the same model, say probit, as that used to predict the decision to purchase, or not, a major household appliance, or automobile, with say, level of income replacing level of radiation dose. Thus, the conditional response of the physician is quantal, or binary, say $\overline{T}/T$, where T is the decision to continue the treatments and $\overline{T}$ is the decision to halt treatments. The "tolerance" of any one radiation oncologist is defined as that critical level of dose, say x_0, which would be just sufficient to elicit the characteristic response, "$\overline{T}$!" The -0.95 quantile of the "tolerance" distribution of the physician, for the event, $\overline{T}/T$, coincides with -0.05 quantile of the "tolerance" distribution of the patient for the event, $\overline{E}_2/E_2$, where E_2 denotes the presence of some dose-limiting complication in normal tissues and $\overline{E}_2$ is the complementary event. (As the ambient levels of litigation - with the imputations of iatrogenic disease - rise the overlap of these two tolerance distributions tends to decrease.) But a more useful definition is provided by a straight-forward analysis of relevant patient responses.

It is, of course, the case that it is only the distribution function, say F(x), of the tolerance, x, that can be observed in dose-response experiments on any system: $P = F(x_0) = {}_{-\infty}\int^{x}0\ f(u)du$, where P is the proportion of individuals "at risk" of a given quantal response for whom the tolerance is $x \leq x_0$. Note that $F(X_0) = P(E|X,Y)$, the conditional probability of the occurrence of the event E, where the exogeneous conditions, e.g., dose, etc., are described by X and the endogenous conditions, e.g., stage of disease, anatomical site, and size, of target volume, etc., are described by Y. N.B. It will be convenient in the remainder of Appendix II to employ an alternative notation for vectors which may now be identified by the letters X or Y, as well as by $\underline{x}^T$, with component identified by x_j. It is not possible to define any characteristic susceptibility of an individual so as to determine by independent observations, the density function, f(x) where $dP = f(x)dx$. Fortunately, in dose-response studies, "... what matters is the dependence of P on dose [and other co-variates] and the unknown parameters and the tolerance distribution is merely a substructure leading to this"(Finney, 1971b). The form of the distribution function is less important, for regression models, than the fact that it provides for a monotonic transformation that maps the finite region of observed response, $0 \leq P \leq 1$, into the infinite response region $-\infty < z < +\infty$, say, if F(x) is Normal and $z = F^{-1}(P)$ is the probit transform of P. The "dose" can, of course, be generalized, in an obvious manner, to an effective dose that includes the several covariates, e.g., fractions, N, elapsed time, T, etc., that modulate the biological effects of a given level of radiation dose - whatever the substantive natures of the binary response and the system (or subsystem) at risk. Then z can be expressed as a linear function (that is, linear in the parameter vector, $\underline{\beta}$) of the dose and other covariates, $z = \underline{x}^T\underline{\beta}$, and the appropriate generalized linear regression methods can be used to obtain the estimates $\underline{\hat{\beta}}$, $Var(\underline{\hat{\beta}})$, RSS, etc. - and the diagnostics (Dobson (1983), Gilchrist (1984), Cook and Weisberg (1982), Pregibon (1981), McCullagh and Nelder (1989)). To repeat, "The important point is that the estimation procedures are not affected by whether the [tolerance] distribution or the formula for P is regarded as fundamental" (Finney, 1971a).

Although there are several quantal responses of irradiated systems in which the assumption of a tolerance distribution is quite implausible, e.g., in radiation-induced mutations, there are others in which it may be quite useful, at least heuristically, to interpret the data in terms of an inherent property of a system - or subsystem - such that if the "dose" is less than a critical level of the property the system will not exhibit the response in question but if the "dose" exceeds this level it will.[a] It seems plausible that the critical level, or tolerance, varies from one individual to another according to a density function, f(x). In such a case, it is also plausible that the critical level for each individual is determined by many independent factors. If these combine additively then the Central Limit Theorem may be used to justify a Normal distribution (Theil, 1971). If these factors combine multiplicatively then the same theorem justifies a Log-Normal distribution (Atchison and Brown, 1966).[b] The tolerance may be a fixed level for any one individual or, more often, it may exhibit both short-term and long-term, or secular, variations in time (if the long-term variations are monotonic the process is described as "aging" - and it may produce either positive or negative increments in the "tolerance" of an individual).

To recapitulate, the tolerance distribution is, "The distribution among a number of individuals of the critical level of a stimulus which will just produce a reaction in each individual.

Although the distribution of these tolerances may be skewed, it is often possible to make it approximately Normal by a simple transformation such as the logarithmic" (Kendall and Buckland, 1971).

Andrews goes on to assert that, "Tolerance has become an operational term ...," (Andrews, 1986) with, apparently, the further implication that this has contributed to some of the difficulties which he proposes to relieve by its deletion from professional discourse. What Andrews intends by the locution, "operational term," is not altogether clear, even from his context. If, however, by "operational" it is meant that the meanings of concepts are derived from, or given by, specific operations, then it seems clear that much of the perceived difficulty with the conventional usage of tolerance in radiation therapy arises from the fact that it has rarely given any concise operational definition. (vide infra).

Still later in his letter Andrews (1986) proposes to reduce the alleged confusion arising from the use of the term "tolerance" - in specialized contexts not anticipated by Webster (1967) - by substitution of the term "risk": "It is just as easy to say that we shall treat to a risk level of 5% as it is to say that we treat to tolerance, but the implications are very different. The former implies that benefit and risk relationships have been taken into account ..." But, in point of fact, Webster (1967) covers very well the risk-benefit context with a precise, operational, definition that is quite different from that implied by Andrews, to wit: "risk ... 4: the product of the amount that may be lost and the probability of losing it." Note here that Webster's "risk" is not dimensionless. This is also the case in any risk-benefit context. Thus, while it may be just as easy to say "... risk level of 5% as it is to say ... tolerance ..." it would not seem that the usage proposed by Andrews is either much more correct - or less confusing. It is, in fact, rather more confusing: Losses must be discussed in terms of what is lost, goods, services, bodily functions, etc. in (usually) monetary equivalents. However, loss ratios are dimensionless (vide infra).

The normative usage, sanctioned by Webster, enjoys a wide currency and may be generalized in an obvious way to take account of the probability distribution of the loss. From econometrics, we find, that "... risk ... is defined as the expected value of loss" (Malinvaud, 1980; Zellner, 1971). This usage is quite similar to that deployed in the classical identification problem (Lachenbruch, 1975; Rao, 1973) and in other enterprises with aleatory, or contingent, outcomes in which incorrect decisions, taken under uncertainty, commonly incur non-trivial penalties. Thus, in Moore and Mendelsohn (1972) cited by Andrews, we find that: "Once costs are assigned, an overall rating of therapy can be estimated by calculating the expected loss, L. This is obtained by summing the products of the cost of each outcome times the probability of its' occurrence ..."

We now present a brief heuristic exposition of normative usages of risk and tolerance in the clinical context, based on response surface models (Myers, 1971) of the concomitant occurrence in the target volume of binary events in two systems - tumor and normal tissue - constructed from actual clinical data (head and neck cancer). These data included observations on the presence/absence of ablation of tumor, the event $\bar{E}_1/E_1$ and necrosis of normal tissue, the event $\bar{E}_2/E_2$, in each target volume. The data are shown in Fig. 1a and 1b. The data for tumor ablation only are described in Fig. 2. The data for normal tissue necrosis have the same distribution of treatment variables - dose and time - but with, of course, different outcomes. So far as we know, this is the only discussion of clinical risk to use "live data" in which the response is a vector, that is, multivariate in the second sense, as discussed in part 6.1.1 of the main body of this report. (N.B.: A more recent bivariate probit model of these data, based on the model of Lesaffre and Molenbergs (1991) and using the software described in their report can be found in Herbert, 1993b.) Figure 3a describes the dose-response surface for the probit model $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ of the data on tumor ablation of Fig. 2. The surface is defined by the grid formed by the families of dose-response curves, $z = \underline{x}^T \underline{\beta}$, and isoeffect curves, $x_1(P) = (z - \beta_0)/\beta_1 - \beta_{2/\beta}1x_2$. Here $x_1 = \log D$, $x_2 = \log T$. There is a cognate surface for the necrosis of normal tissue in the target volume. The level of uncomplicated control achieved at a given treatment regimen depends on the relative positions of these two plane surfaces. Figure 3b describes a family of dose-response curves, $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \beta_0^* + \beta_1 x_1$, for the data of Fig. 2. The slope, β_1, is invariant and the

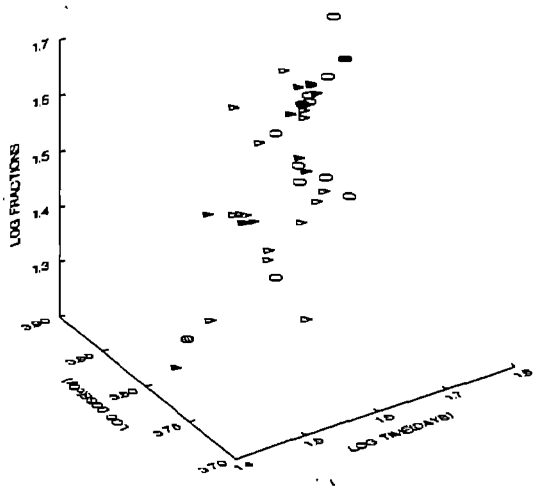


Fig. 1a. 3-dimensional scattergram of the joint binary radiation responses in tumor and normal tissues for n=45 patients with head and neck cancer. Circles - tumour recurrence; triangles - tumor ablation. Filled symbols - normal tissues necrosis; open symbols - no necrosis. The cross-hatched circle identifies that observation for which the regression diagnostic, Cook's D, for the bivariate probit model of these data is largest (Lesaffre and Molenberghs, 1991. See Herbert, 1993b).

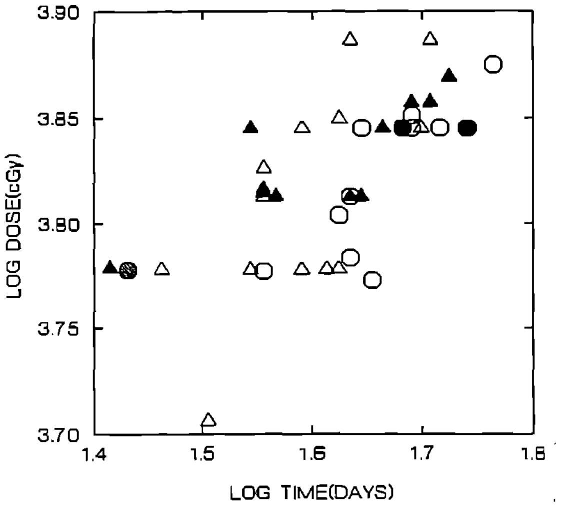


Fig. 1b. Scattergram of n=45 observations on the radiation response of patients with head and neck cancer. The symbols are as in Fig. 1a.

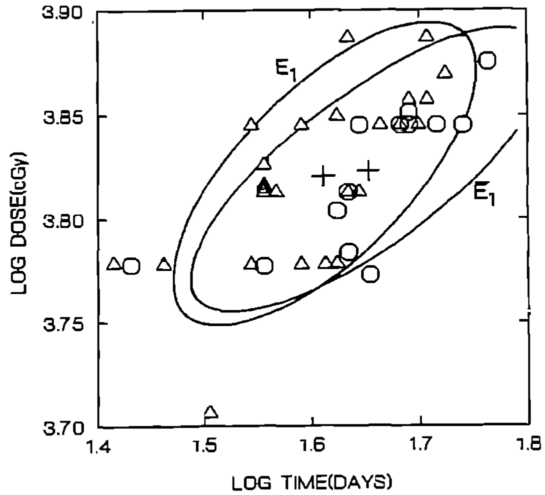


Fig. 2. The figure presents a superposition of the respective 0.80 contour ellipses (Hald, 1952) on the scattergrams of the responses E_1 (ablation) and $\bar{E}_1$ (recurrence) in 46 patients treated for head and neck cancer. The Mahalanobis distance, D_m^2, is a measure of the overlap of the distributions of E_1 and $\bar{E}_1$ and hence of the degree to which a useful model of a dose-response surface can be constructed. For useful models to be constructed the overlap must be small: $D_m^2 > 3$-4. For these data $D_m^2 < 1.0$. D_m^2 is the sample estimate of $\Delta^2 = (\mu(E_1) - \mu(\bar{E}_1)T\Sigma^{-1}(\mu(E_1) - \mu(\bar{E}_1))$ where $\mu(E_1)$ and $\mu(\bar{E}_1)$ are the mean vectors of the respective conditional distributions of the treatment regimens (x_1, x_2) and Σ is the common covariance matrix of the two distributions. D^2 is a biased estimate of Δ^2: $D_m^2 > \Delta^2$.

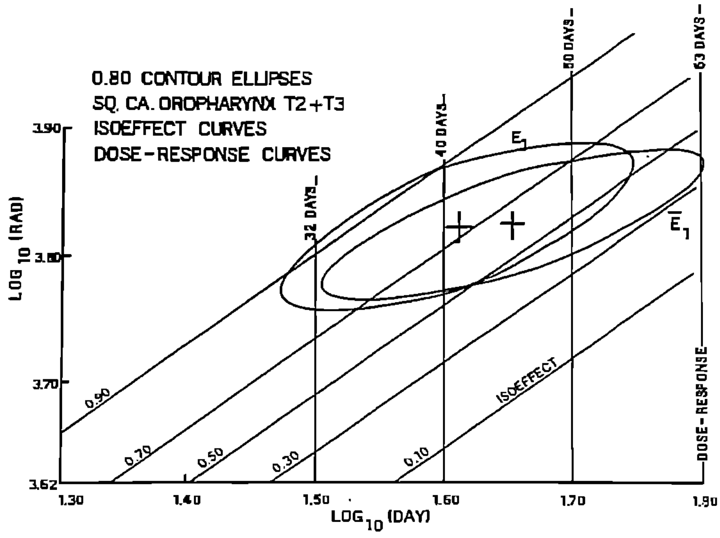


Fig. 3a. The figure presents a superposition of the probit response surface for the model, $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ of the data in Fig. 2 and the respective contour ellipses for the event E_1 and $\bar{E}_1$. The surface is defined by the families of dose-response and isoeffect curves. Because the overlap of the respective contour ellipses $\bar{E}_1$ and E_1 is so large for these data, they are of largely heuristic value.

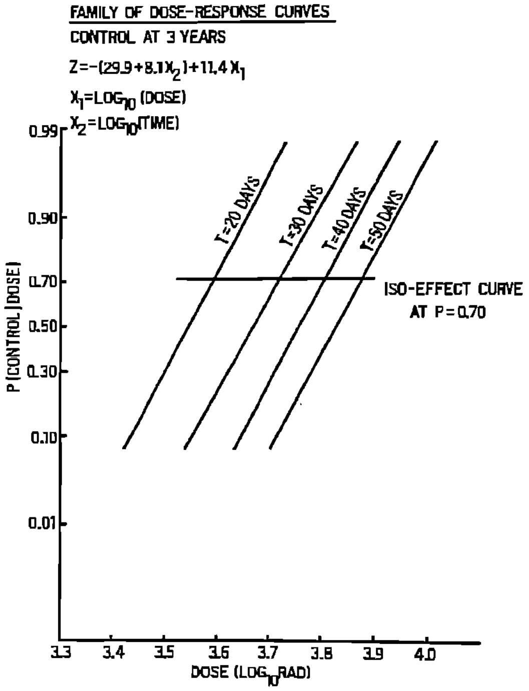


Fig. 3b. The figure presents a family of dose-response curves for the probit model of the data of Fig. 2 for several different levels of the time T.

intercept, β_0^*, is a function of T.

Figure 4a presents two alternative measures of the relative positions at a fixed value of T of the respective response surfaces for ablation and for necrosis. One measure is the familiar therapeutic ratio, $R = x_1^N(0.05) - x_1^T(0.95)$. $x_1^N(0.05)$ is the 0.05 quantile of the distribution of the response, necrosis. $x_1^T(0.95)$ is the 0.95 quantile of the distribution of the response, ablation.

The therapeutic ratio, R, is a measure of the relative position of the two response surfaces in a direction parallel to the dose - axis. The second measure, $C = P(E_1) - P(E_2) = P_1 - P_2$, is a measure of the relative position in a direction parallel to the response - axis. A brief recapitulation of elementary probability theory seems in order: If $\bar{E}_1/E_1$ and $\bar{E}_2/E_2$ are the binary events, in the target volume, ablation/recurrence (tumor) and necrosis/no necrosis (normal tissue), respectively, then, obviously, $P(E_1 \text{ and } E_2) + P(E_1 \text{ and } \bar{E}_2) + P(\bar{E}_1 \text{ and } E_2) + P(\bar{E}_1 \text{ and } \bar{E}_2) = 1.0$. The joint event of clinical interest is the binary event, $\bar{S}/S$: treatment success, $S = E_1 \text{ and } \bar{E}_2$ and treatment failure, $\bar{S} = \bar{E}_1 \text{ or } E_2$. Evidently $P(\bar{S}) + P(S) = 1.0$. It is clear that, $P(\bar{S}) = P(E_1 \text{ and } E_2) + P(\bar{E}_1 \text{ and } E_2) + P(\bar{E}_1 \text{ and } \bar{E}_2)$. If $\bar{E}_1$ and E_2 are mutually exclusive events, i.e., the probability of recurrence of tumor at levels of dose high enough to necrose normal tissues is quite small - which is often quite reasonable - then $P(\bar{E}_1 \text{ and } E_2) = 0$. Thus, $P(\bar{S}) = P(\bar{E}_1) + P(E_2)$ and $P(S) = 1-P(\bar{S}) = 1-P(\bar{E}_1) - P(E_2) = P(E_1) - P(E_2)$. See Table 1. In our data (see Fig. 2) the event $\bar{E}_1$ and E_2 did not occur. (The joint event, $\bar{E}_1$ and E_2, is, of course, Paterson's "supralethal dose effect" - about which one does not hear much anymore.) Professor Cohen (Cohen, 1960) has correctly pointed out that if the events E_1 and $\bar{E}_2$ are mutually independent then the probability of uncomplicated control, the event E_1 and $\bar{E}_2$, is given by $P(S) = P(E_1)*P(\bar{E}_2)$, not $P(S) = P(E_1 - P(E_2)$. He has also remarked that if $P(E_2) > P(E_1)$ then $P(\bar{S}) < 0$. Both remarks are, of course, quite true. However, it must be noted that it seems unlikely that the events, E_1 and E_2, are independent if both occur in the target volume, as is the case we have considered here - the data of Table 1b describe complete association (Fienberg, 1980). Moreover, if $P(E_2) > P(E_1)$ where, again, E_1 and E_2 are events that occur within the target volume, this suggests that the tumor is highly radioresistant - or the treatment plan is inadequate. See section 6.1 for a more general treatment of the joint radiation effects. (N.B.: The synthesis of the information on the occurrence of the joint event, $\bar{E}_1$, and concomitant necrosis, E_2, into the binary variable, Treatment Success, $S = E_1 \text{ and } \bar{E}_2$, and its complement, Treatment Failure, $\bar{S} = \bar{E}_1 \text{ or } E_2$, is both of interest and of use. However, currently and historically, treatments are seldom deliberately carried to a point at which the yield of necrosis exceeds a few percent. Rather, only a slightly larger yield of a less serious concomitant complication, E_2, such as fibrosis, determines the intensity of an irradiation schedule, i.e., fibrosis is the treatment-limiting complication. For these complications, the assumption, $P(\bar{E}_1 \text{ and } E_2) \equiv 0$, is no longer true and consequently the simple representation of the yield of treatment success in terms of the marginal probabilities, $P(S) - P(E_1) - P(E_2)$, is no longer valid. Instead, we must estimate the yield, $P(S|X)$, directly as the joint probability, $P(S|X) = P(E_1 \text{ and } \bar{E}_2|X)$. See Herbert 1993b for an example.)

It is evident that for data in which $P(\bar{E} \text{ and } E_2) = 0$, $C = P_1 - P_2$ in Fig. 4a is a measure of the yield of the event, S, Treatment Success: $P(S) \equiv P(E_1 \text{ and } \bar{E}_2)$. Figure 4b describes the relation between the dose-response curves for $P(\bar{E}_1|X)$, $P(E_2|X)$, and $P(S|X)$, for target volumes and treatment regimens, such that $P(\bar{E}_1 \text{ and } E_2) = 0$, in which the yield of success is $P(S|X) = P(E_1|X) - P(E_2|X)$.

Figure 4c is the family of dose-response curves for the joint event $S = E_1 \text{ and } \bar{E}_2$ and thus is cognate to Fig. 4b. Figure 4d describes a family of isoeffect curves for the yield of the joint effect, concomitant control of tumor and absence of complication in normal tissue:

$P(E_1 \text{ and } \bar{E}_2|X) = P(E_1|X) - P(E_2|X) = P_1 - P_2 = C.$

The position of the mean treatment regimen, $\bar{x}_1$, $\bar{x}_2$, of Fig. 2 is at X. The joint yield of this regimen is 0.39. A possible local maximum in the joint yield is at the regimen Z where the joint yield is 0.61. (N.B.: Figure 3b, 4b, 4c, and 4d are reproduced, with permission, from Herbert, 1978.)

The surface shown in Fig. 4e provides a classical example of a feature which commonly

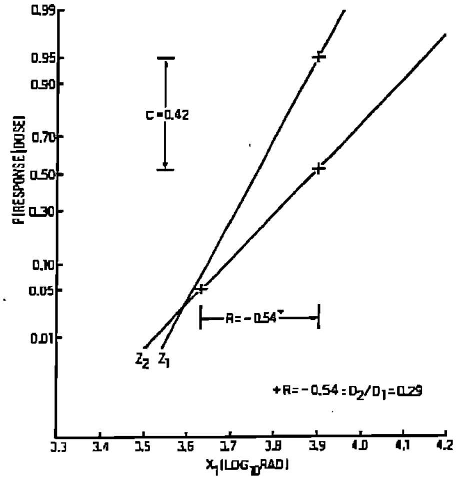


Fig. 4a. The figure presents a superposition of the respective dose-response curves z_1 and z_2 for the events E_1 and E_2, control and necrosis at three years, respectively. The therapeutic ratio, say $R(0.05,0.95) = \log_{10}D_2 - D_1$, and the probability of uncomplicated control, say $C = P_1 - P_2$, for protraction T = 40 days are also shown.

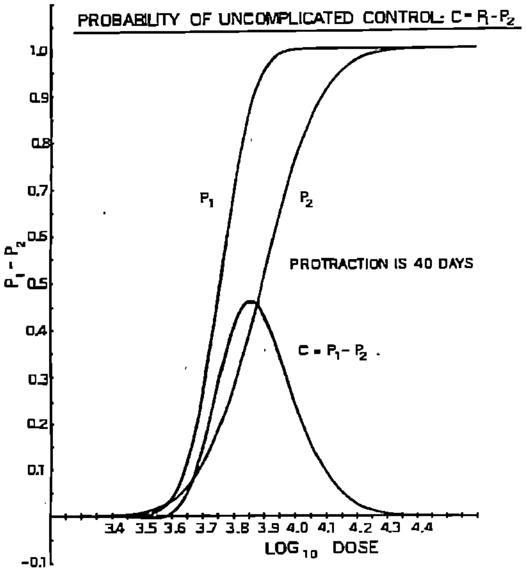


Fig. 4b. The figure presents a superposition of the respective dose-response curves for the events E_1, E_2, and $S = E_1$ and $\bar{E}_2$.

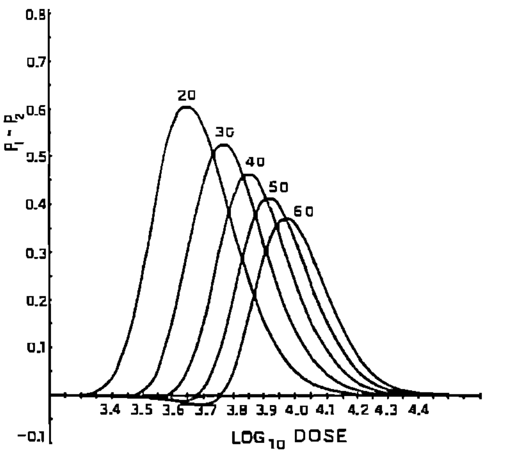


Fig. 4c. The figure presents a family of dose-response curves for the event S for several different values of T.

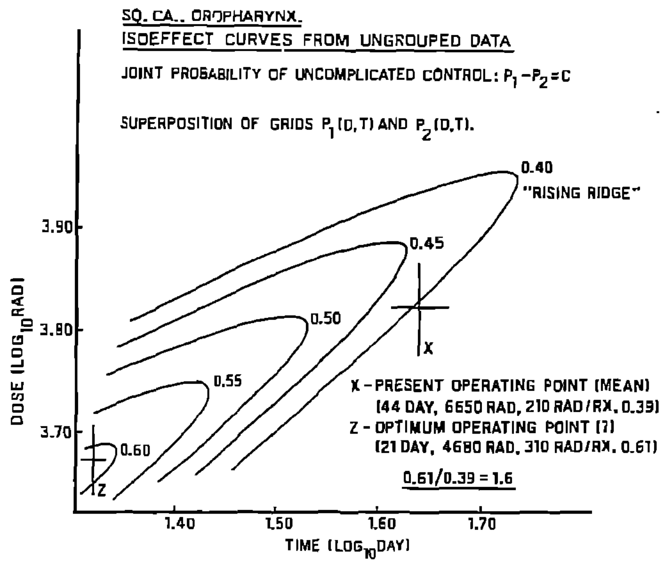


Fig. 4d. The figure presents a family of isoeffect curves for the event S for several different levels of P. The point "X" describes the location of the mean, ($\bar{x}_1$, $\bar{x}_2$), of the treatment regimens of Fig. 2.

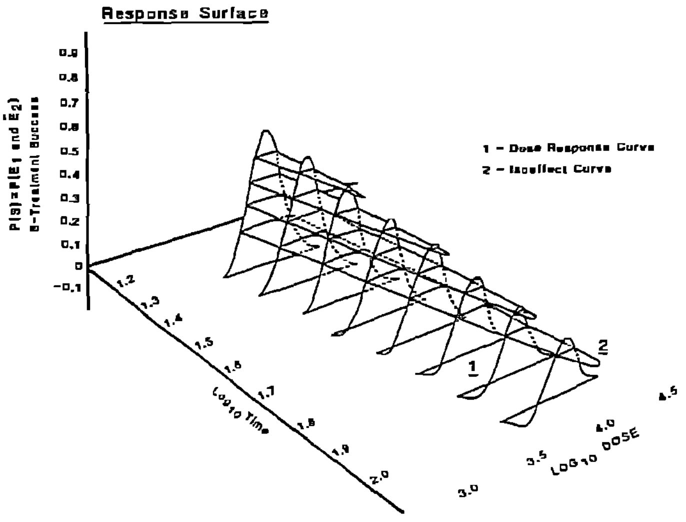


Fig. 4e. The figure presents a perspective of the response surface for the event S.

Table __1a__. Multivariate response in target volume.

		Ablation	Recurrence	
		E_1	$\bar{E}_1$	
Necrosis	E_2	$P(E_1$ __and__ $E_2)$	$P(\bar{E}_1$ __and__ $E_2)$	$P(E_2)$
No Necrosis	$\bar{E}_2$	$P(E_1$ __and__ $E_2)$	$P(\bar{E}_1$ __and__ $E_2)$	$P(\bar{E}_2)$
		$P(E_1)$	$P(\bar{E}_1)$	1.0

Binary events:

1. E_1 - Ablation of tumor
   $P(\bar{E}_1) + P(E_1) = 1.0$

2. E_2 - Necrosis of normal tissue
   $P(\bar{E}_2) + P(E_2) = 1.0$

3. S - Treatment success
   $P(\bar{S}) + P(S) = 1.0$
   $\bar{S} = \bar{E}_1$ __or__ E_2
   $S = E_1$ __and__ $\bar{E}_2$

$P(E_1 \text{ and } \bar{E}_2) + P(E_1 \text{ and } E_2) + P(\bar{E}_1 \text{ and } E_2) + P(\bar{E}_1 \text{ and } \bar{E}_2) = 1.0.$
$P(\bar{S}) = P(E_1 \text{ and } E_2) + P(E_1 \text{ and } E_2) + P(E_1 \text{ and } E_2)$
$P(S) = P(E_1 \text{ and } \bar{E}_2)$
$P(\bar{S}) + P(S) = 1.0$

Table __1b__. Joint Response in Target Volume. Sq. Ca. Oropharynx.

		Tumor Ablation, E_1	Recurrence, $\bar{E}_1$	Totals
Normal	Necrosis E_2	14	0	14
Tissue	No Necrosis $\bar{E}_2$	17	14	31
	Totals	31	14	45

The joint event on uncomplicated control is defined as a treatment success; it is the event S. Then,

$$P(S) = P(E_1 \text{ and } \bar{E}_2) = 1 - P(E_1 \text{ and } E_2) - P(\bar{E}_1 \text{ and } E_2) - P(\bar{E}_1 \text{ and } \bar{E}_2)$$

But, since $P(\bar{E}_1 \text{ and } E_2) = 0$, then

$P(S) = 1 - P(E_1 \text{ and } E_2) - P(\bar{E}_1 \text{ and } \bar{E}_2) = 1 - P(E_2) - P(\bar{E}_1) = P(E_1) - P(E_2)$

occurs in response surface analyses: This feature, a "rising ridge" is characteristic of the yield or response surfaces of many chemical processes during the early stages of operation. (Davies, 1967 and Myers, 1971) The surface itself is a geometric representation of the outcome of the "pattern of care" which prevailed at this institution during the period in which these patients described in Fig. 2 were irradiated. Figure 4e presents a 3-dimensional perspective of the dose-response surface for the joint event S. The surface is defined by the grid formed by the families of dose-response (Fig. 4c) and isoeffect (Fig. 4d) curves for the joint event S.

We next define risk in terms of the losses, L_1, L_2, respectively, incurred by the patient upon the occurrence in the target volume of either of the binary events, $\bar{E}_1$, recurrence of disease, or E_2, necrosis in the target volume (which is taken to be the dose-limiting complication) with the respective probabilities $(1-P_1)$ and P_2. The risk is, therefore, $L = L_1(1-P_1) + L_2P_2$. Figure 5a is a plot of the dose-response curves $(1-P_1)$ and P_2 for the events $\bar{E}_1$ and E_2, respectively, where $P_j = P(E_j|X,Y)$, $j = 1,2$ and the risk curve $L = L_1(1-P_1) + L_2P_2$ for $L_1 = L_2 = 1$. Figure 5b is the plot of a family of risk curves.

This operational concept of risk provides still another operational definition of tolerance dose: For a fixed value of x_2, the minimum risk occurs at the radiation dose, x_1^*, where $(\partial L/\partial x_1) = -L_1(\partial P_1\partial x_1) + L_2(\partial P_2/\partial x_1) = 0$. It is useful to describe the dose-response curve for the event $\bar{S} = \bar{E}_1$ or E_2, that is, $P(\bar{S}|X,Y)$, as a risk curve. The value of dose, x_1^*, for which the risk is a minimum is now defined to be the tolerance dose for that value of x_2. It denotes the dose to which that 'reasonable and prudent' radiation oncologist may carry the irradiation (for a given value of time, x_2) since at this dose the probability of failure in either mode is minimal (but different from zero).

Figure 5c presents a 3-dimensional perspective of a risk surface - the response surface for the event, treatment failure, $\bar{S} = (\bar{E}_1$ or $E_2)$. The tolerance curve is the locus of the extrema of the risk curves for $L_1 = L_2 = 1$. It is clear that both the position and slope of the tolerance curve in the x_1-x_2 plane are functions of the loss ratio, L_1/L_2, assessed by the oncologist.[c] Therefore, a tolerance curve is not an isoeffect curve, since the respective probabilities, $P(\bar{S})$, $P(\bar{E}_1)$ and $P(E_2)$ all vary along the curve. It is instead an iso-constraint curve $[(\partial L/\partial x_1)_x \equiv 0]$. It does not appear that this distinction has been made before, although it is clearly not without clinical significance.

Figure 5d is a superposition of two families of curves: 1) Risk curves: $L = L_1(1-P_1) + L_2P_2$, 2) Tolerance curves: $x_1^* = C_0 + C_1x_2$. The risk curves are shown for values of elapsed time $T = 25$, 40 and 63 days. The tolerance curves are shown for values of the loss ratio $L_1/L_2 = 0.5$, 1.0 and 2.0. Risk curves describe the variation of loss to the patient at treatment failure (in either mode, recurrence or necrosis) as the dose, x_1, is varied at fixed time, x_2. However, tolerance curves describe the variation of the "tolerance dose" x_1^* (the local minimum of risk) as the time, x_2, is varied. The location and shape of the tolerance curve is determined, in part, by the loss ratio, L_2/L_1. (N.B.: Figure 5b and 5d are reproduced, with permission, from Herbert, 1978.) The response surface described in Fig. 5c appears to represent an apt mathematicization of Wither's recent remark: "There is no such thing as a tolerance dose for a tissue or organ, only for a tissue or organ in a given set of clinical circumstances. The tolerance dose has been defined as the regimen that produces the maximum acceptable probability of complications in a given treatment situation" (Withers, 1989). Here, the maximum acceptable probability of occurrence of a complication, the event E_2 is determined by the probability of the concomitant occurrence of the event E_1, tumor ablation. More generally, however, the response surfaces for the occurrence of the events $\bar{S}$ and S provide a mathematical description of the realization to the object of radiation therapy: "The object of clinical radiation therapy is to obtain, for each patient, the maximum probability of cancer cure while minimizing the likelihood of significant normal tissue damage" (R. Yaes, 1988).

Thus, the tolerance curve, $x_1^* = \alpha_0 + \alpha_1x_2$ is simply the locus of these points (x_1, x_2) at which the risk is an extremum: $(\partial L/\partial x_1) = 0$. For the head and neck data with $L_1 = L_2 = 1.0$ we have, $x_1^* = 2.764 + 0.681x_2$, or, $D^* = 575T^{0.681}$. For $T = 40$ day, $N = 30$, we have $D^* = 7160$ cGy, yielding $P(E_1) = 0.86$, $P(E_2) = 0.40$ and $P(S) = P(E_1) - P(E_2) = 0.46$. In the set of data to which we had access the overall levels of $P(E_1)$ and $P(E_2)$ were 0.65 and 0.30, respectively; P(S)

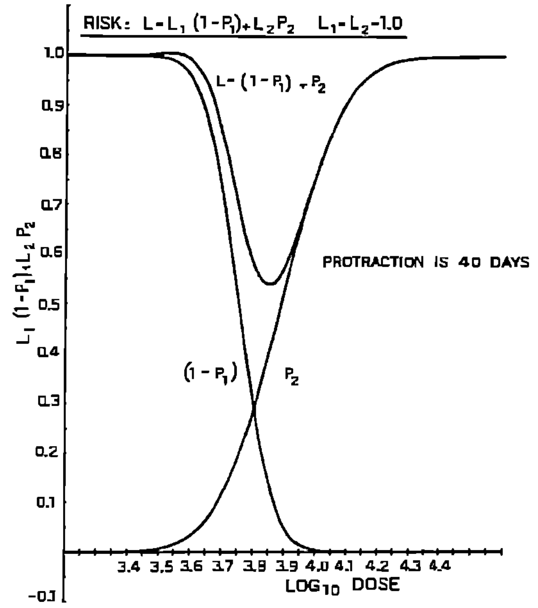


Fig. 5a. The figure presents a superposition of dose-response curves for the events E_1 and E_2 and the risk curve L, for $L_1 = L_2 = 1$.

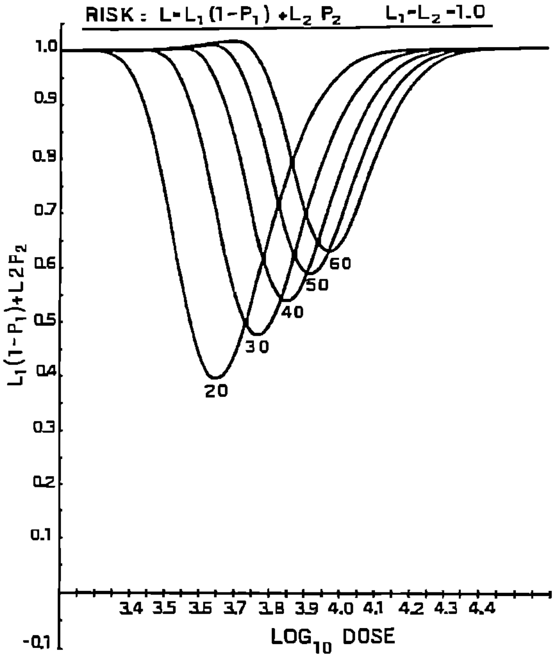


Fig. 5b. The figure presents a family of risk curves L for several different levels of T.

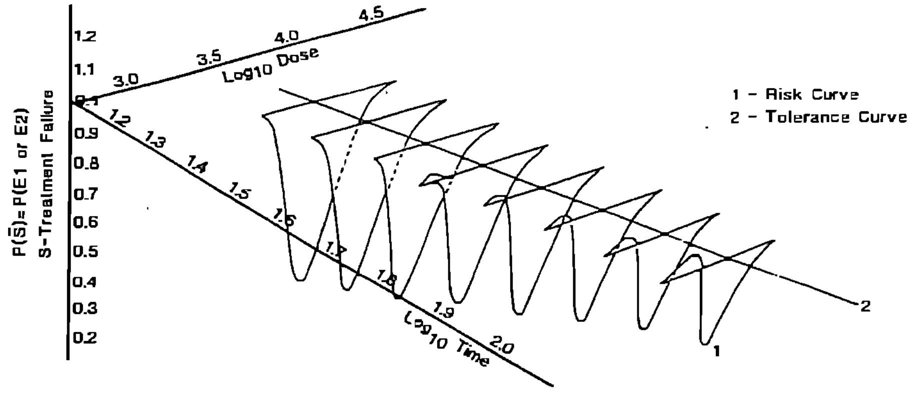


Fig. 5c. The figure presents a perspective view of the family of risk curves of Fig. 5b and the tolerance curve which is the locus of the extrema of the respective risk curves.

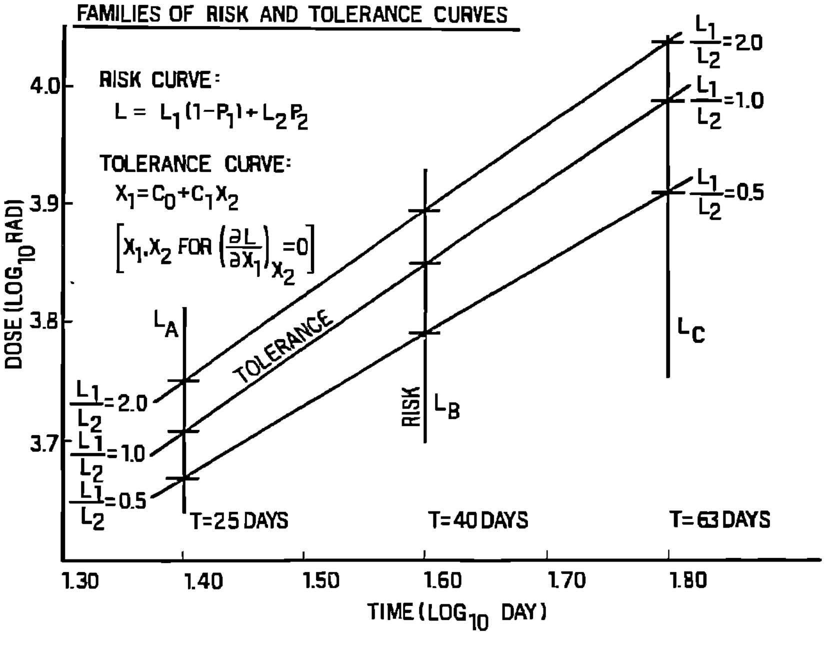


Fig. 5d. The figure presents the superposition of a family of risk curves for different levels of T and tolerance curve for different loss ratios. If the radiation therapist can specify a loss ratio, the tolerance curve is determined by the data. This methodology, "lets the data speak" (Ehrenberg, 1976) to the issue of tolerance.

= 0.35. (Too often, even "the best of a bad situation" is not so very good.[d])

It is quite evident that unless there is an explicit quantitative relation between the yields, $P(E_i|X, Y)$, of each of the several sequelae (control, complication) and the adjustable variables of treatment, X, for a patient of specified prognostic features, Y, the aim of "treating to tolerance" must remain a goal which is more suitable for admiration than to regular achievement. It is this explicit, quantitative representation of this relation which the response surface methods we have described has provided. But we can take the response surface concept still further:

1. The effect of prognostic features.

The prognostic vector Y is implicit in the yield equation which is explicit in the treatment vector X since the parameter vector, $\underline{\beta}$, is a function of the prognostic vector, Y. That is $\underline{\beta} = \underline{\beta}(Y)$. This method of accounting for the effects of the prognostic features of the patient upon the outcomes to an irradiation schedule is quite analogous to several of the current and highly instructive methods of including the effects of such features upon survival ($\bar{E}/E$ - Dead/Alive) following treatment of other diseases by other maneuvers. In these later studies, a function of the hazard, or age-specific death-rate, λ, for the empirical exponential survival curves ("time-response curves") is represented as a linear form on the prognostic features of the patient.

In the classic paper of Myers et al, 1966, the hazard function, λ - the parameter of the time-response (survival) curve for breast cancer treated by surgery - can be represented as the linear regression on four dichotomized prognostic features, y_k:[e] $-\ln\lambda = \alpha_0 + \alpha_1 y_1 + \alpha_2 y_2$ ..., where $Y = (y_1, \ldots y_4)$. Four prognostic features, each at two levels, define $2^4 = 16$ distinct prognostic strata, each with its characteristic survival curve: $S_j = \exp(\lambda_j t)$, $1 \leq j \leq 2^4 = 16$. By analogy the parameters, β_j, of the dose-response surface, can be represented as $\beta_j = \alpha_0 + \alpha_1 y_1 + \ldots$, j=0, ..., p; $\underline{\beta}^T = (\beta_0, \beta_1 \ldots \beta_p)$.

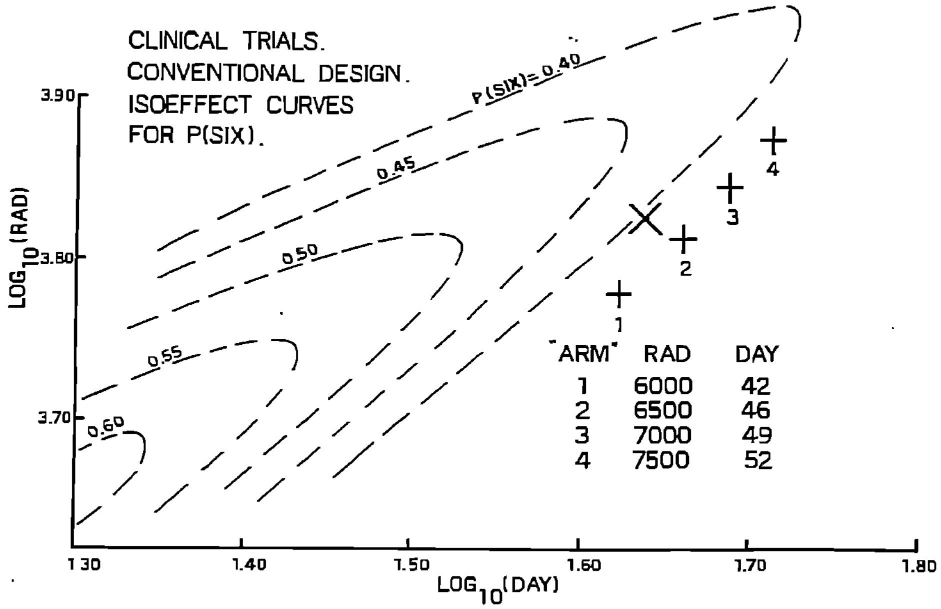
More recently, Schultheiss and Orton have presented several important, as well as interesting, alternative developments of this important topic using statistical decision theory (Schultheiss and Orton 1984; Schultheiss, 1981).

2. Experimental designs for response surfaces.

> "The objective is to maximize the probability of tumor control without incurring an unacceptably high probability of organ damage."
>
> R. Yaes, 1989

Conventional clinical trials are designed to test a hypothesis (on Treatment A vs Treatment B) not to estimate the parameter vector of dose-response models. (See part 8.1 of main body of the report and part 3 of Annex II for discussion of designs for parameter estimation.) The design matrix of a hypothetical clinical trial (conventional or eu-fractionation) with four "arms" is shown in Fig. 6a. It is super-posed on the family of isoeffect curves for the event S - which are assumed to be unknown to the designer of the trial. (Compare with Fig. 4d). It is evident that there will be "no significant difference" in the respective levels of S achieved since the set of treatment regimens lies nearly parallel to the family of isoeffect (a constant level of S) curves in that region. (We remark that this seems to be the most common result: Most clinical trials in chronic and malignant disease disclose that "there is no significant difference between treatments A and B." The frequent negative results may often be due to lack of any true difference - as well as to the small numbers of patients at risk discussed in section 6.1.4 of the present report that gives rise to a high false negative error rate.)

But, the ultimate aim of all clinical trials is, of course, to provide neither an efficient test of hypotheses on treatment nor efficient estimates of the form and parameters of a yield equation but rather to find the treatment regimen, X^*, for which the yield of treatment success is a maximum for patients within a specified prognostic stratum. The search should be an evolutionary procedure: the results of one trial should inform the design of subsequent trials and, thereby, a sequence of trials is generated in steps which lead to the optimum treatment. This evolution is readily implemented with the use of response surface methods to design the distribution of treatment regimens and to assess the local shape of the surface from data generated from the design



"ARM"	RAD	DAY
1	6000	42
2	6500	46
3	7000	49
4	7500	52

Fig. 6a. The figure presents a superposition of the design matrix of a conventional clinical trial on the family of isoeffect curves for the event S. It is apparent that there will be no significant difference in the response between the regimens because the set of treatment regimens lies parallel to the isoeffect curve. The point "X" is the location of the mean, $(\bar{x}_1, \bar{x}_2)$, of the treatment regimens of Fig. 2.

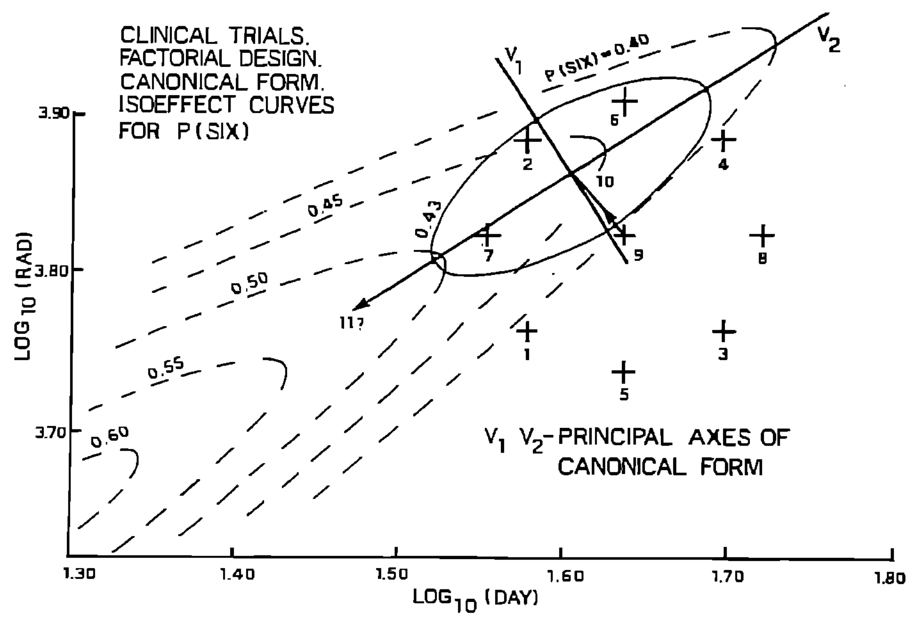


Fig. 6b. The figure presents a superposition of the treatment matrix of the central composite rotatable factorial designs of a clinical trial on the family of isoeffect curves for the event S. It can be shown that this can be incorporated into a sequential method that will find and follow "ridges" in the second-order response surface for S to the location of a maximum response, i.e., a regimen at which the probability of uncomplicated control is a global maximum. Note that the analysis of the experimental data, leads quite easily from point 9, the current "operating point", to point 10, a local maximum of S, to point 11 at which P(S|X) is still greater, and so on. Point 9 describes the location of the mean, $(\bar{x}_1, \bar{x}_2)$, of the treatment regimens of Fig. 2. The experimental design is described in Table 2.

at each step of the sequence. An important part of the problem in such sequential designs is to determine the smallest number of experiments that will find the regimen X at which the yield P(S|X) is a maximum.

We now present a simulation that will illustrate the usefulness of the response surface methodology. We assume that the response surfaces that we have constructed previously from the data of Fig. 2 (and the cognate data on necrosis in the target volume) and described in Figs. 3-5 are unknown to us and must be discovered by experiment. However, the outcomes of these subsequent experiments will be estimated from these models.

The nine points of the design described in Table 2 and Fig. 6b will permit the estimation of the parameters of a second order surface in the region of the design. The event of interest is the joint event, treatment success, S, rather than simple events of tissue status, such as ablation, E_1, or necrosis, E_2: S = E_1 and $\bar{E}_2$.

There are four possible systems of isoeffect (constant P(S|X)) curves for a yield equation of second degree in two control variables: a) a maximum, b) a stationary ridge, c) a rising ridge and d) a minimax or col (a saddle point). The equation to be derived now is one which relates the conditional probability of success, P(S|X), itself to the variables of treatment. It is a multiple regression of yield, y = P(S|X), on the treatment variables:

$$y = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_{12} x_1 x_2 + \delta_{11} x_1^2 + \delta_{22} x_2^2.$$

It will be useful to rewrite this equation in matrix form:

$$y = \alpha + \underline{x}^T \underline{\beta} + 0.5 \underline{x}^T \Gamma \underline{x}.$$

$\underline{\beta}$ is the (2*1) vector of linear coefficients (δ_1, δ_2) and Γ is the (2*2) symmetric matrix in which the diagonal elements are twice the quadratic terms ($2\delta_{11}$, $2\delta_{22}$) and the off-diagonal elements are the interaction terms (δ_{12}). The canonical form of y is expressed in terms of the eigenvalues (λ_1, λ_2) and eigenvectors (V_1, V_2) of the matrix Γ of coefficients of terms of second degree in x_1, x_2: $y = \sigma + \lambda_1 V_1^2 + \lambda_2 V_2^2$. Here σ is the level of y at the stationary point $\underline{X}^o = -\Gamma^{-1}\underline{\beta}/2$. The signs and sizes of the eigenvalues (λ_1, λ_2) identify the local nature of the surface as one of the four types a) - d) described above. For example, for a local maximum, we have $\lambda_1 < 0$, $\lambda_2 < 0$. (For a local minimum, $\lambda_1 > 0$, $\lambda_2 > 0$.) If λ_1 and λ_2 differ in sign then the surface is, locally, a saddle or col. If $\lambda_1 < 0$, $\lambda_2 < 0$ with $|\lambda_2| << |\lambda_1|$ and the stationary point is at the edge of, or beyond, the region of the data, then the surface is, locally, a rising ridge with the direction of rise given by the eigenvector V_2. There is an abundant statistical literature on the sequential design of factorial experiments (in particular the so-called central composite rotatable designs) that will identify the local nature of the response surface, e.g., a local maximum, or a rising ridge, and will suggest the location and direction of subsequent experiments to "follow" the latter to a global maximum (or minimum). See Box and Wilson (1951), Davies (1967), and Myers (1971). See also Herbert (1978). The parameter vector $\underline{\delta}$ is an implicit function of the prognostic features of the patient just as in the case of the parameter vector of the equations for the probit transforms, z_i. However, the parameters for the equation for the event S are estimated by the method of Least Squares, whereas the method of Maximum Likelihood is used to estimate the parameters for the equations for the events E_1 and E_2.

We have approximated the probability of success, P(S|X), by a second degree Taylor series in the region of the space of treatment variables that is centered on the mean of the set of treatment regimens, X = (x_1, x_2), of a preceding retrospective study. In the present example, the analysis of the data of Fig. 2 provides this location. At each of the 9 treatment regimens of the central composite rotatable design, X_j, of Fig. 6b and Table 2a the probability of treatment success, S, is estimated by the difference in the respective probabilities of occurrence of each event, E_1 and E_2. See Table 2b.

The application of Least Squares methods to these data gives the parameter estimates for the equation,

$$P(S|X) = -137.4 + 83.2x_1 - 28.0x_2 + 13.2x_1 x_2 - 13.5x_1^2 - 7.3x_2^2.$$

The first degree terms, δ_1, δ_2, are considerably larger than those of second degree which suggests that stationary point in P(S|X) does not lie within the region of the design. It is necessary to

Table 2a. Central Composite Rotatable Design in Two Variables (Myers, 1971). Coded Variables, x_i^*.

#	x_0	x_1^*	x_2^*	x_1^{*2}	x_2^{*2}	$x_1^* x_2^*$
1)	1	-1	-1	1	1	1
2)	1	1	-1	1	1	-1
3)	1	-1	1	1	1	-1
4)	1	1	1	1	1	1
5)	1	$-\sqrt{2}$	0	2	0	0
6)	1	$\sqrt{2}$	0	2	0	0
7)	1	0	$-\sqrt{2}$	0	2	0
8)	1	0	$\sqrt{2}$	0	2	0
9)	1	0	0	0	0	0

$x_i^* = 2(x_i - \bar{x}_i)/d_i$. $d_i = x_i(\text{max}) - x_i(\text{min})$.
See Fig. 6b.

N.B. For k independent variables of treatment, X_1, X_2, ..., X_k, a first order design requires 2^k treatment regimens. To estimate terms of second order requires an additional 2k + 1 treatment regimens or $2^k + 2k + 1$ regimens in all. Since at least $n_i = 30$ patients must be allocated to each regimen it is clear that the number, k, of treatment variables to be contemplated for any yield equation cannot exceed three or four.

Table 2b. Design Matrices for a 2^2 Central Composite Rotatable Design.

| ← 2nd Order Matrix → |

| ← 1st Order Matrix → |

Point, i.	1	2	3	4	5	6	7	8	9
	Coordinates (Transformed and Coded Variables)								
x_1	-1	1	-1	1	$-\sqrt{2}$	$\sqrt{2}$	0	0	0
x_2	-1	-1	1	1	0	0	$-\sqrt{2}$	$\sqrt{2}$	0
	Coordinates (Natural Variables)								
D(cGy)	5794	7638	5794	7638	5470	8091	6653	6653	6653
T(day)	38	38	50	50	44	44	36	53	44
$P(E_1\|X)$	0.60	0.95	0.24	0.74	0.30	0.92	0.87	0.41	0.68
$P(E_2\|X)$	0.23	0.50	0.13	0.36	0.13	0.49	0.34	0.20	0.28
$P(S\|X)$	0.37	0.45	0.11	0.38	0.17	0.43	0.53	0.21	0.40

Statistics of Distribution of X = (D, T)

$\bar{D}$ = 6709, ΔD = 2621. $\bar{T}$ = 44, ΔT = 17.0. r_{DT} = 0.

estimate the location of the stationary points. These are the regimens $X^0 = (x_1^0, x_2^0)$ for which

$$(\partial P(S|X)/\partial x_1) = (\partial P(S|X)/\partial x_2) = 0.$$

It is readily shown that for the present set of data the treatment regimen, $X^0 = (3.86, 1.61) = (7280\ \text{cGy}, 40\ \text{day})$, is at an extremum of $P(S|X)$. At this regimen the yield is $P(S|X^0) = 0.46$.

The extremum can, of course, be a maximum, a minimum, or a col - a saddle - of the response surface. These may be distinguished from one another by construction of the canonical form of the polynomial yield equation. (Davies, 1967 and Myers, 1971) This is the equation

$$P(S|X) = P(S|X^0) + \lambda_1 V_1^2 + \lambda_2 V_2^2.$$

The λ_1, λ_2 and V_1, V_2 are the eigenvalues and eigenvectors, respectively, of the quadratic form of the second order Taylor series approximation to $P(S|X)$:

$$\lambda_1 = -17.70, \ V_1^T = (0.845, -0.534)$$

$$\lambda_2 = -3.11, \ V_2^T = (0.534, 0.845)$$

Since $\lambda_1 < 0$ and $\lambda_2 < 0$, the extremum at $X^0 = (3.86, 1.61)$, is a (local) maximum. The concomitant yields of ablation and necrosis at this regimen are, respectively, $P(E_1|X^0) = 0.87$, $P(E_2|X^0) = 0.41$.

The set of treatment regimens which comprise the design of the experiment is shown in Fig. 6b. The contour ellipse $P(S|X) = 0.43$ of the canonical form for the quadratic approximation of $P(S|X)$ is also shown. The contour is, of course, just the 0.43 isoeffect curve for the joint event, treatment success, $S = E_1$ and E_2. The "true" isoeffect curves (dashed) for $P(S|X) = 0.40$, 0.45, 0.50, 0.55 and 0.60 are superimposed upon the experimental design and the canonical contour.

It will be immediately noticed that the "true" isoeffect curves and the one constructed upon the designed experiment are both conic sections: the former are hyperbolae, the latter an ellipse. It will also be noticed that the concordance of the two estimates of the shape of the surface in the region in which they overlap is "not too bad". However, the small difference is most important. The family of isoeffect curves (dashed) for $P(S|X)$ describe the "true" response surface for which the experiment was intended to provide a local description. These curves define a rising ridge which has a global maximum at some position in the lower left quadrant of the graph. At this maximum, $P(S|X) - 0.60$. The experiment, however, has described a local maximum to be at the origin $(x_1^0, x_2^0) = (0, 0)$ of the principal axes, V_1, V_2. The apparent discrepancy can be readily resolved. It was noted previously that the first order terms in the polynomial representation of $P(S|X)$ are dominant. In itself this suggests that the maximum may be remote from the region of the design. It was also remarked that the origin of the principal axes was at the periphery of the data. This simply suggests that the canonical form of the equation may not accurately describe the (local) shape of the surface in the region of the design.

A better appreciation of the nature of the response surface in the region of the design can be obtained by rewriting the canonical equation in a principal axis system, $V_1^\#$, $V_2^\#$, for which the origin is that of the experimental design. The coordinates of the latter in the original (V_1, V_2) system are (-0.05, + 0.007). Therefore, the origin of the new set of principal axes, $(V_1^\#, V_2^\#)$, is taken to be at the point (-0.05, + 0.007) in the (V_1, V_2) system. In the new system the canonical equation has the form

$$P(S|X) = 0.416 - 17.70V_1^{\#2} - 3.11V_2^{\#2} + 1.77V_1^\# - 0.04V_2^\#.$$

The presence of the first order terms show that the initial design was in the vicinity of a ridge of the response surface. This is quite consistent with Fig. 6b. The slope of the ridge in the direction of the $V_{1\#}$ axis is 1.77. The slope of the ridge in the direction of the $V_2^\#$ axis is -0.04, that is, the location of the treatment regimen, X_1^{00} at which the yield is a global maximum lies along the ridge, "up and to the left," with respect to the center of the design.

Subsequent experiments will be necessary to more accurately describe the position of the optimum treatment, say X^{00}. The center of the design for the next experiment should be at the origin (0,0) of the first principal axis system $V_1^\#$, $V_2^\#$. The pairs of design points $(\sqrt{2}, 0)$, $(-\sqrt{2}, 0)$ and $(0, \sqrt{2})$, $(0, -\sqrt{2})$ should be on the $V_1^\#$ and $V_2^\#$ axes, respectively.

The figure shows that the procedure which has just been described is an evolutionary one. More technically, it is a procedure for recognizing and following the ridges in the response surface which lead from local to global maxima. The experimental study has moved systematically from the treatment regimen at point 9, (x_1, x_2) = (6653 rad, 44 day), at which the yields are $P(S|X) = 0.40$, $P(E_1|X) = 0.68$, $P(E_2|X) = 0.28$ to the regimen at point 10, (x_1^0, x_2^0) = (7280 rad, 40 day), at which the yields are $P(S|X^0) = 0.46$, $P(E_1|X^0) = 0.87$, $P(E_2|X^0) = 0.41$. Equally important, the study has characterized the nature of the response surface in the region of the design. The next experiment will move the study along the ridge toward the global maximum of $P(S|X)$, perhaps in the region of point 11.

The virtues of the response surface methods are quite impressive when compared with those of the current alternative, the conventional linear design of experiments. Obviously, there is no important difference in $P(S|X)$ at any of the arms of one such design that is shown in Fig. 6a superimposed upon the family of true isoeffect curves for $P(S|X)$. One of the difficulties with the conventional design is immediately clear: for this disease, the line of the design lies parallel to the isoeffect curve at $P(S|X) \equiv 0.34$. Furthermore, it can be seen that subsequent experiments, for which the design is informed by the received practice, will describe (ambiguously) a weak local maximum as the sequence of experiments crosses the ridge system at a point where $P(S|X)$ ¯ 0.38. At this point the evolutionary sequence will "stick". The conventional procedures of design and evaluation cannot recognize and subsequently follow the ridges in the response surface of $P(S|X)$. They will, therefore, seldom lead to optimum treatment regimens.

It is of interest to note that the results of a recent clinical trial are consistent with this diagnosis of a major and inherent defect in the conventional design of clinical trials: the several alternative treatment regimens frequently lie along an isoeffect curve in $x_1 - x_2$ space. The treatment regimens and the respective yields for the recent trial are described in Table 3. (As the trial serves only a heuristic purpose we have not identified it further.)

We have shown that the conventional design of clinical trials may be a rather remarkably crude procedure by which to assess and inform those therapeutic maneuvers which are directed at producing local effects in the patient. For an expenditure of a reasonable number of patients, very little can be learned and that little, not very well. The conventional design is inadequate for a test of hypotheses upon the yields of alternative treatments because the regimens lie upon the isoeffect curves of the joint event S. It will not provide any information on the local shape of the surface and hence of the direction and distance in the data space from the region of the present design to another region, X_j, at which the yield may be greater, i.e., it can neither recognize nor follow a ridge in the response surface.

While the conventional design is inefficient in the sense that for the expenditure of a given number of patients the state of the medical art is advanced very little, it seems to be ethical in the sense that the risk of either type of failure for any given patient is held to values which do not differ much between regimens of the design since the scale of this design is small. Therefore, no single patient will be irradiated at a regimen for which the yield of ablation (and hence the yield of concomitant necrosis) is very much greater than that of his fellows or for which the yield of recurrence is very much greater than theirs. However, for a given precision of estimates, the size, n, of the study and the shape and scale of the distribution of treatments, vary inversely with one another. The small scale of the conventional design requires that a large number, n, of patients be at risk for each type of binary failure in the trial. Clearly, the scale and shape of the study, as well as its size, have ethical as well as statistical features which must be taken into account in the design.

It is possible that the clinical inequities inherent in good experimental design can be ameliorated to some degree by combination therapy, i.e., those patients that are allocated to regimens for which the probability of recurrence is unacceptably high could be included in a subsequent chemotherapy trial. Similarly, those for whom the probability of necrosis is higher could be candidates for reconstructive surgery. On the other hand, the weak information that could be obtained from a trial that included less extreme regimens that educe more acceptable levels of recurrence and necrosis could perhaps be strengthened by Bayesian methods using information obtained in isomorphic animal experiments (See part 14.2 of the main body of this report).

Table 3. Design and outcomes for an actual clinical trial.

	4000#	4000 rad	5000	6000
Regimen (dose in rad)	(split)	in 4 wks.	in 5 wks.	in 6 wks.
Total regression (E_1^*)	7%	23%	22%	21%
Total plus Partial Regression (E_1)	47%	50%	58%	57%
Severe Complications(E_2)	15%	10%	7%	17%
100P (S)	32	40	51	40##

# 2000/1 wk + 0/2 wks + 2000/1 wk.

## This suggests the presence of a "ridge" in the surface of P(S). See Fig. 6a.

We remind the reader that the foregoing development is solely heuristic: We have presented a rival point of view on the design of clinical trials, and the methodology by which it may be implemented. It can be generalized, in obvious ways, to accommodate the more usual circumstance in which the correlation structures of the tolerance and treatment vectors are more complex, in particular, in which the multivariate response consists of both acute and chronic radiation effects in normal tissue and acute effects in tumor and the treatment must be specified by measures of total dose, D, fractionation, N, and protraction, T.

Footnotes

[a] As a general remark, radiation-induced responses are observed either as counts or as proportions. It is only for the latter that the concept of a tolerance distribution may be useful.

[b] The empirical argument that the "dose" must be positive will also lead to the log Normal distribution of tolerance.

[c] x_1 = logD, x^2 = logT (For this heuristic development it is assumed that the number of fractions is N=(5/7)T since that is the case for these data.)

[d] We caution that these are small sample estimates and thus have motivational value only. Wholesome discussions of the difficulties encountered in obtaining estimates of $P(E_1)$, $P(E_2)$ from either experimental or non-experimental data may be found in Annexes II-IV.

[e] These prognostic features are:

Axillary lymph nodes, y_1 =	-1	involved
	+1	not involved
Nuclear grade, y_2 =	-1	undifferentiated
	+1	differentiated
Sinus histocytosis, y_3 =	-1	absent
	+1	present
Tumor size, y_4 =	-1	≥ 5 cm
	+1	< 5 cm

REFERENCES

AAAS/ABA (1988) Project on Scientific Fraud and Misconduct. Report on Workshop Number One. National Conf. of Lawyers and Scientists. Hedgesville, WV. Sept. 18-20, 1987. AAAS. Washington, DC.

AAAS/ABA (1989) Project on Scientific Fraud and Misconduct. Report on Workshop Number Two. National Conf. of Lawyers and Scientists. Queenstown, MD. Sept. 23-25, 1988. AAAS. Washington, DC.

AAAS/ABA (1989) Project on Scientific Fraud and Misconduct. Report on Workshop Number Three. National Conf. of Lawyers and Scientists. Irvine, CA. Feb. 17-18, 1989. AAAS. Washington, DC.

Acton, F.S. (1959) Analysis of Straight-Line Data. Dover Pub. NY.

Agren, A., Brahme, A. and Turesson, I. (1990) Optimization of Uncomplicated Control for Head and Neck Tumors. Intl. J. Radia. Oncol. Biol. Phys. 19: 1077-1085.

Aitchison, J. and Brown, J.A.C. (1966) The Lognormal Distribution. Univ. Press. Cambridge.

Aitchison, J. and Silvey, S.D. (1957) The Generalization of Probit Analysis to the Case of Multiple Responses. Biometrika. 44: 131-140.

Akaike, H. (1974) A New Look at the Statistical Model Identification. IEEE Trans. on Automatic Control. Ac-19(6): 716-723.

Akaike, H. (1977) On Entropy Maximization Principle. IN Applications of Statistics. 27-41. P.R. Krishaiah, ed. North-Holland Pub. NY.

Akaike, H. (1985) Prediction and Entropy. IN A Celebration of Statistics. 1-24. A.C. Atkinson & S.E. Fienberg, eds. Springer-Verlag. NY.

Allen, D.M. (1971) The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables. Technical Report No. 23. Dept. of Statis.Univ. of Ky.

Allen, D.M. (1974) The Relationship Between Variable Selection and Data. Argumentation and a Method for Prediction. Technometrics. 16(1): 125-127.

Allen, D.M. and Jordan, D.C. (1982) The Use of Prior Information for Prediction. Biometrics. 38: 787-799.

Alper, T. (1980) Keynote Address: Survival Curve Models. IN Radiation Biology in Cancer Research. 3-18. R.E. Myers and H. R. Withers, eds. Raven Press. NY.

Altman, D.G. (1982) Statistics and Ethics in Medical Research. IN Statistics in Practice. 1-21. Br. Med. Assn. London.

Ames, B.N., Magaw, R. and Gold, L.S. (1987) Ranking Possible Carcinogenic Hazards. Science. 236: 271-280.

Anderberg, M.R. (1973) Cluster Analysis for Applications. Academic Press. NY.

Anderson, P.W. (1972) More is Different. Science. 177. 393-396.

Andrews, J.R. (1982) Optimization of Radiotherapy: Some Notes on the Principles and Practice of Optimization in Cancer Treatment and Implications for Clinical Research. Cancer Clin. Trials. 4: 483-495.

Andrews, J.R. (1986) Tolerance: An Anachronism. Int. J. Radia. Oncol. Biol. Phys. 12: 289-293.

Ang, K.K., van der Kogel, A.J. and van der Schueren, E. (1983) The Effect of Small Radiation Doses on the Rat Spinal Cord: The Concept of Partial Tolerance. Int. J. Radia. Oncol. Biol. Phys. 9: 1487-1491.

Anscombe, F.J. (1961) Examination of Residuals. 4th Berkeley Symposium on Mathematical Statistics and Probability. 1-36. 4th ed. J. Neyman, ed. Univ. of Calif. Press. Berkeley & LA.

Armitage, P. (1971) Statistical Methods in Medical Research. Blackwell Scientific Publications. Oxford.

Ashford, J.R. and Sowden, R.R. (1970) Multi-Variate Probit Analysis. Biometrics. 26: 535-546.

Aydelotte, W.O. (1966) Quantification in History. The Amer. Hist. Rev. LXXI: 803-825.

Babbage, C. (1830/1971) Reflections on the Decline of Science in England. Irish Univ. Press. Shannon, Ireland.

Bacon, F. (1620/1960) The New Organon. Bobbs-Merrill Co. Inc. NY.

Bacon, Roger (1271/1962) The Opus Majus of Roger Bacon. Vol. 1. Russell & Russell, Inc. NY.

Bailar, J.C. (1982) Research Quality, Methodologic Rigor, Citation Counts, and Impact. AJPH. 72(10): 1103-1104.

Bailar, J.C. and Mosteller, F., eds. (1986) Medical Uses of Statistics. NEJM Books. Waltham, MA.

Bailar, J.C. and Smith, E.M. (1986) Progress Against Cancer? New Eng. J. Med. 314: 1226-1232.

Bailey, N.T.J. (1967) The Mathematical Approach to Biology and Medicine. John Wiley & Sons. NY.

Baker, R.J. and Nelder, J.A. (1978) The GLIM System. Release 3: Generalized Linear Interactive Modelling Manual. Numerical Algorithms Group. Oxford.

Barber, B. (1961) Resistance by Scientists to Scientific Discovery. Science. 134: 596-602.

Bard, Y. (1974) Nonlinear Parameter Estimation. Academic Press. NY.

Barendsen, G.W. (1978) Fundamental Aspects of Cancer Induction in Relation to the Effectiveness of Small Doses of Radiation. IN: Late Biological Effects of Ionizing Radiation. V. II. 263-275. IAEA. Vienna.

Barendsen, G.W. (1982) Dose Fractionation, Dose Rate and Isoeffect Relationships for Normal Tissue Responses. Int. J. Radia. Oncol. Biol. Phys. 8: 1981-1997.

Barton, M.B., Keane, T.J., Gadalla, T. and Maki, E. (1992) The Effect of Treatment Time Interruption on Tumour Control Following Radical Radiotherapy of Cancer. Radiotherapy & Oncol. 23: 137-143.

Bates, D.M. and Watts, D.G. (1988) Nonlinear Regression Analysis and Its Applications. John Wiley & Sons. NY.

Beaton, A.E., Rubin, D.B. and Barone, J.L. (1976) The Acceptability of Regression Solutions: Another Look at Computational Accuracy. J. Amer. Statis. Assn. 71: 158-168.

Begg, C.B. (1985) A Measure to Aid in the Interpretation of Published Clinical Trials. Statis. in Med. 4: 1-9.

Belsley, D. A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons. NY.

Bentzen, S.M., Christensen, J.J., Overgaard, J. and Overgaard, M. (1987) Some Methodological Problems in Estimating Radiobiological Parameters From Clinical Data. Acta Oncologica. 27: Fasc. 2. 105-116.

Bentzen, S.M. and Thames, D.H. (1991) Clinical Evidence for Tumor Clonogen Regeneration: Interpretations of the Data. Radiother. and Oncol. 22: 161-166.

Berger, J. (1982) Bayesian Robustness and the Stein Effect. J. Amer. Statis. Assn. 77: 358-368.

Berkson, J. (1953) A Statistically Precise and Relatively Simple Method of Estimating the Bioassay with Quantal Response, Based on the Logistic Function. J. Amer. Statis. Assn. 48: 565-599.

Berlin, J.A., Begg, C.B. and Louis, T.A. (1989) An Assessment of Publication Bias Using a Sample of Published Clinical Trials. J. Amer. Statis. Assn. 84: 381-392.

Berlin, J.A., Laird, N.M., Sacks, H.S. and Chalmers, T.C. (1989) A Comparison of Statistical Methods for Combining Event Rates from Clinical Trials. Statis. in Med. 8: 141-151.

Berry, M.V. (1976) Waves as Catastrophes. Physics Bulletin. 26(1):

Black, B. (1988a) A Unified Theory of Scientific Evidence. Fordham Law Rev. 56: 595-695.

Black, B. (1988b) Evolving Legal Standards for the Admissibility of Scientific Evidence. Science. 239: 1508-1512.

Blakemore, et al (1964) Source not further identified.

Blumberg, M.S. (1986) Risk Adjusting Health Care Outcomes: A Methodologic Review. Medical Care Rev. 43: 351-393.

Boag, J.Q., Osborn, S.B. and Rotblat, J. (1959) Criteria for a Threshold in Radiation Effects. Br. J. Radiol. 32: 70.

Bode, Mostellar, Tukey, and Winsor (1949) Source not further identified.

Bolles, R. C. (1988) Why You Should Avoid Statistics? Biol. Psychiatry. 23:

Bond, V.P., Cronkite, E.P., Lippincott, S.W. and Shellabarger, C.J. (1960) Studies on Radiation-Induced Mammary Gland Neoplasia in the Rat. III. Relation of the Neoplastic Response to Dose of Total-Body Radiation. Radia. Res. 12: 276-285.

Box, G.E.P. (1960) Fitting Empirical Data. Annals New York Acad. Sci. 86(3): 792-816.

Box, G.E.P. (1979) Robustness in the Strategy of Scientific Model Building. In: Robustness in Statistics. 201-236. R.L. Launer and G.N. Wilkinson, eds. Academic Press. NY.

Box, G.E.P. (1980) Sampling and Bayes' Inference in Scientific Modelling and Robustness (with discussion). J. Royal Statis. Soc. A: 143: Pt. 4. 383-430.

Box, D.E.P. (1984) The Importance of Practice in the Development of Statistics. Technometrics. 26: 1-8.

Box, G.E.P. and Draper, N.R. (1969) Evolutionary Operation. John Wiley & Sons. NY.

Box, G.E.P. and Draper, N.R. (1987) Empirical Model-Building and Response Surfaces. John Wiley & Sons. NY.

Box, G.E.P. and Hunter, W.G. (1965) The Experimental study of Physical Mechanisms. Technometrics. 7: 23-42.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) Statistics for Experimenters. John Wiley & Sons. NY.

Box, G.E.P. and Lucas, H.L. (1959) Design of Experiments in Non-Linear Situations. Biometrika. 46: 77-90. 1959.

Box, G.E.P. and Tiao, G.C. (1973) Bayesian Inference in Statistical Analysis. Addison-Wesley Pub. Co. Reading, MA.

Box, G.E.P. and Wetz, J. (1973) Criteria for Judging Adequacy of Estimation by an Approximating Response Function. Tech. Report #9. Dept. of Statis. Univ. Wisconsin. Madison.

Box, G.E.P. and Wilson, K.B. (1951) On the Experimental Attainment of Optimum COnditions. J. Royal Statis. Soc. B. 13:

Box, M.J. (1971) Bias in Nonlinear Estimation (with discussion). J. Royal Statis. Soc. B. 33(2): 171-201.

Breslow, N.E. and Day, N.E. (1980) Statistical Methods in Cancer Research. Vol. 1. The Analysis of Case-Control Studies. IARC Scientific Publ. No. 32. Lyon.

Breslow, N.E. and Storer, B.E. (1985) General Relative Risk Functions for Case- Control Studies. Amer. J. Epidem. 122: 149-162.

Bross, I.D.J. (1970) The Role of Mathematical Models in Clinical Research. Amer. Statistician. 24(1): 53-56.

Brown, C.C. (1975) On the Use of Indicator Variables for Studying the Time-Dependence of Parameters in a Response-Time Model. Biometrics. 31: 863-872.

Brown, R.G. (1962) Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice-Hall Inc. Englewood Cliffs, NJ.

Bruer, J.T. (1982) Methodological Rigor and Citation Frequency in Patient Compliance Literature. AJPH. 72(10): 1119-1123.

Buck, C. (1975) Popper's Philosophy for Epidemiologists. Intl. J. Epidem. 4(3): 159-168.

Bunge, M. (1979) Causality and Modern Science. 3rd ed. Dover Pub. NY.

Bury, K.V. (1975) Statistical Models in Applied Science. John Wiley. NY.

Carroll, J.B. (1961) The Nature of the Data, or How to Choose a Correlation Coefficient. Psychometrika. 26(4): 347-371.

Carroll, R.J. and Ruppert, D. (1984) Power Transformations When Fitting Theoretical Models to Data. J. Amer. Statis. Assn. 79: 321-328.

Carroll, R.J. and Ruppert, D. (1988) Transformation and Weighting in Regression. Chapman and Hall. NY.

Casella, G. (1985) An Introduction to Empirical Bayes Data Analysis. Amer. Stat. 39: 83-87.

Casti, J.L. (1989) Alternate Realities. Mathematical Models of Nature and Man. John Wiley. NY.

Chalmers, T.C. (1990) The Quality of Primary and Secondary Research and Meta-Analysis. Annals of Theoretical Surgery. 14: 323.

Chalmers, T.C., Levin, H., Sacks, H.S., Reitman, D., Berrier, J. and Nagalingam, R. (1987) Meta-Analysis of Clinical Trials as a Scientific Discipline. I: Control of Bias and Comparison with Large Co-Operative Trials. Statis. in Med. 6: 315-325.

Chalmers, T.C., Berrier, J., Sacks, H.S. Levin, H., Reitman, D. and Nagalingam, R. (1987) Meta-Analysis of Clinical Trials as a Scientific Discipline. II: Replicate Variability and Comparison of Studies that Agree and Disagree. Statis. in Med. 6: 733-744.

Chapman, J.D. (1980) Biophysical Models of Mammalian Cell Inactivation by Radiation. IN Radiation Biology in Cancer Research. 21-32. R.E. Meyn & H.R. Withers, Raven Press. NY.

Chatterjee, S. and Hadi, A.S. (1988) Sensitivity Analysis in Linear Regression. John Wiley & Sons. NY.

Chatterjee, S. and Price, B. (1977) Regression Analysis by Example. John Wiley & Sons. NY.

Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. J. Amer. Statis. Assn. 74: 829-836.

Cochran, W.G. (1973) Experiments for Nonlinear Functions. J. Amer. Statis. Assn. 68: 771-781.

Cohen, L. (1960) The Statistical Prognosis in Radiation Therapy. A Study of Optimal Dosage in Relation to Physical and Biologic Parameters for Epidermoid Cancer. Amer. J. Roentgen. 84: 741-753.

Cohen, L. (1966) Radiation Response and Recovery: Radiobiological Principles and Their Relation to Clinical Practice. IN The Biological Basis of Radiation Therapy. Schwartz, E.E., ed. 208-348. J.B. Lippincott Co. Philadelphia, PA.

Cohen, L. (1982) The Tissue Volume Factor in Radiation Oncology. Intl. J. Radia. Oncol. Biol. Phys. 8: 1711-1774.

Cohen, L. (1983) Biophysical Models in Radiation Oncology. CRC Press. Boca Raton, FL.

Cohen, L. (1987) Optimization of Dose-Time Factors for a Tumor and Multiple Associated Noral Tissues. Intl. J. Radia. Oncol. Biol. Phys. 13: 251-258.

Cohen, M.R. and Nagel, E. (1934) An Introduction to Logic and Scientific Method. Harcourt, Brace & World, Inc. NY.

Collins, R. (1987) Comment on, "Meta-Analysis of Clinical Trials as a Scientific Discipline: I. Control of Bias and Comparison with Large Cooperative Trials," by T.C. Chalmers, H. Levin, H. Sacks, D.Teitman, J. Berrier and R. Nagalingam. Statis. in Med. 6: 327.

Conniffe, D. and Stone, J. (1973) A Critical View of Ridge Regression. The Statis. 22(2): 181-187.

Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman & Hall. NY.

Cook. T.D. (1974) The Potential and Limitations of Secondary Evaluations. Educational Evaluation: Analysis and Responsibility. 155-222. M.W. Apple, et al, eds. McCatchan Pub. Berkeley, CA.

Cornfield, J. (1962) Joint Dependence of Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: a discriminant function analysis. Fedn. Proceedings. 21: 58-61.

Couch, N.P., Tilney, N.L., and Rayner, A.A. (1981) The High Cost of Low-Frequency Events. NEJM. 304: 634-637.

Courant, R. and John, F. (1965) Introduction to Calculus and Analysis. Vol. 1. Wiley & Sons. NY.

Coveney, P. and Highfield, R. (1990) The Arrow of Time. Fawcett Columbine. NY.

Cox, C. (1990) Fieller's Theorem, the Likelihood and the Delta Method. Biometrics. 46: 709-718.

Cox, D.R. (1970) Analysis of Binary Data. Methuen & Co. Ltd. London.

Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. Chapman & Hall. NY.

Cox, D.R. and Snell, E.J. (1968) A General Definition of Residuals (with discussion). J. Royal Statis. Soc. B: 30: 248-275.

Cramer, E.M. (1964) Some Comparisons of Methods of Fitting the Dosage Response Curve for Small Sample. Amer. Statis. Assn. Journal. 779-793.

Cramer, E.M. (1974) Brief Report: The use of Highly Correlated Predictors in Regression Analysis. Multivariate Behavioral Research. 9(2): 241-243.

Dale, R. G. (1985) The Application of the Linear-Quadratic Dose-Effect Equation to Fractionated and Protracted Radiotherapy. Br. J. Radiol. 58: 515-528.

Daniel, C. and Wood, F.S. (1971) Fitting Equations to Data. John Wiley & Sons. NY.

Darby, S.C. (1986) Some Recent Statistical Analyses of Two Long-term Studies of Exposure to Ionizing Radiation. Statis. in Med. 5: 539-546.

Darlington, R.C. (1978) Reduced-Variance Regression. Psychological Bulletin. 85(6): 1238-1255.

Davies, O.L., ed. (1961) Statistical Methods in Research and Production. Hafner Pub. Co. NY.

Davies, O.L., ed. (1967) The Design and Analysis of Industrial Experiments. Hafner Pub. Co. NY.

Dawkins, R. (1986) The Blind Watchmaker. Norton. NY.

Dempster, A.R. (1983) Purposes and Limitations of Data Analysis. Scientific Inference, Data Analysis, and Robustness. 117-133. Academic Press. NY.

Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977) A Simulation Study of Alternatives to Ordinary Least Squares (with comments). J. Amer. Statis. Assn. 72: 77-106.

DerSimonian, R., Charette, L.J., McPeek, B. and Mosteller, F. (1982) Reporting on Methods in Clinical Trials. NEJM. 306: 1332-1337.

Descartes, R. (1955) The Philosophical Works of Descartes. Trans. E.S. Haldane and G. Ross. Dover Pub. NY.

Diaconis, P. (1985) Theories of Data Analysis: From Magical Thinking Through Classical Statistics. IN Exploring Data Tables, Trends, and Shapes. 1-36. D.C. Hoaglin, F. Mosteller and J.W. Tukey, eds. John Wiley & Sons. NY.

Diderrich, G.T. (1985) The Kalman Filter From the Perspective of Goldberger-Theil Estimators. Amer. Statis. 39(3): 193-198.

Dische, S. (1991a) A Review of Hypoxic Cell Radiosensitization. Int. J. Radia. Oncol. Biol. Phys. 20: 147-152.

Dische, S. (1991b) Advances in Basic Science: Have they Benefited Patients with Cancer? The Brit. J. Radiol. 64: 1081-1091.

Dobson, A.J. (1983) Introduction to Statistical Modelling. Chapman and Hall. NY.

Dolby, G.R. (1982) The Role of Statistics in the Methodology of the Life Sciences. Biometrics. 38: 1069-1083.

Donabedian, A. (1978) The quality of medical care. Science. 200: 856-864.

Douglas, B.G. and Fowler, J.F. (1976) The Effect of Multiple Small Doses of X-Rays on Skin Reactions in the Mouse and a Basic Interpretation. Radia. Res. 66: 401-426.

Draper, N.R. and John J.A. (1988) Response-Surface Designs for Quantitative and Qualitative Variables. Technometrics. 30(4): 423-428.

Draper, N.R. and Smith, H. (1981) Applied Regression Analysis. 2nd ed. John Wiley & Sons. NY.

DuMouchel, W.H. (1990a) ASA Conference on Radiation and Health, VIII. 107-111. Proceedings of the ASA Conf. 9-13, 1989. Copper Mountain, CO.

DuMouchel, W.H. (1990b) Bayesian Meta-analysis. IN Statistical Methodology in the Pharmaceutical Sciences. D. Berry, ed. Marcel Dekker, NY.

DuMouchel, W.H. and Groer, P.G. (1989) A Bayesian Methodology for Scaling Radiation Studies from Animals to Man. Health Phys. 57: 411-418.

DuMouchel, W.H. and Harris, J.E. (1983) Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species. J. Amer. Statis. Assn. 78: 293-315.

DuMouchel, W.H. and Herbert, D.E. (1993) Combining Information from Multiple Dose/Time/Fractionation Studies (Bayesian Hierarchical Meta-Analysis Revisited). IN Prediction of Response in Radiation Therapy: Radiosensitivity and Repopulation. (In Press) Proceedings of the 4th Intl. Conf. on Dose, Time, and Fractionation in Radiation Oncology. Madison, WI. Sept. 1992.

Dykstra, O. (1971) The Augmentation of Experimental Data to Maximize $[X^TX]$. Technometrics. 12(3): 682-688.

Easterbrook, P.J., Berlin, J.A., Gopalan, R. and Matthews, D.R. (1991) Publication Bias in Clinical Research. Lancet. 337: 867-872.

Eddington, A. (1959) New Pathways in Science. Cambridge Univ. Press.

Efron, B. (1975) The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. J. Amer. Statis. Assn. 70: 892-898.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans. SIAM. Philadelphia, PA.

Efron, B. (1986) How Biased is the Apparent Error Rate of a Prediction Rule? J. Amer. Statis. Assn. 81: 461-470.

Efron, B. and Morris, C. (1973) Stein's Estimation Rule and Its Competitors - An Empirical Bayes Approach. J. Amer. Statis. Assn. 68: 117-130.

Ehrenberg, A.S.C. (1968) The Elements of Lawlike Relationships. J. Royal Stat. Soc. Series A. 131: 280-301.

Ehrenberg, A.S.C. (1975) Data Reduction. John Wiley & Sons. NY.

Ellis, F. (1969) Dose, Time and Fractionation: A Clinical Hypothesis. Clinical Radiol. 20(1): 1-7.

Ellwood, P.M. (1988) Shattuck Lecture - Outcomes Management. A Technology of Patient Experience. NEJM. 318: 1549-1556.

Feinstein, A.R. (1990) Scientific News and Epidemiologic Editorials: A Reply to the Critics. Epidem. 1(2): 170-180.

Feinstein, A.R. and Horwitz, R.I. Double Standards, Scientific Methods, and Epidemiologic Research. NEJM. 307: 1611-1617.

Fertil, B., Dertinger, H., Courdi, A. and Malaise, E.P. (1984) Mean Inactivation Dose: A Useful Concept for Intercomparison of Human Cell Survival Curves. Radia. Res. 99: 73-84.

Fertil, B., Deschavanne, P.J., Gueulette, J., Possoz, A., Wambersie, A. and Malaise, E.P. (1982) In Vitro Radiosensitivity of Six Human Cell Lines. II. Relation to the RBE of 50-MeV Neutrons. Radia. Res. 90: 526.537.

Fertil, B., Deschavanne, P.J., Lachet, B. and Malaise, E.P. (1980) In Vitro Radiosensitivity of Six Human Cell Lines. Radia. Res. 82: 297-309.

Fertil, B. and Malaise, E.P. (1981) Inherent Cellular Radiosensitivity as a Basic Concept for Human Tumor Radiotherapy. Int. J. Radia. Oncol. Biol. Phys. 7: 621-629.

Fertil, B. and Malaise, E. (1985) Intrinsic Radiosensitivity of Human Cell Lines is Correlated with Radioresponsiveness of Human Tumors. Analysis of 101 Published Survival Curves. Int. J. Radia. Oncol. Biol. Phys. 11: 1699-1707.

Field, S.B., Hornsey, S., and Kutsutani, Y. (1976) Effects of Fractionated Irradiation on Mouse Lung and a Phenomenon of Slow Repair. Br. J. Radiol. 49: 700-707.

Fienberg, S.E. (1980) The Analysis of Cross-Classified Categorical Data. 2nd ed. MIP Press. Cambridge, MA.

Fienberg, S.E. (1989) Source not further identified.

Filliben, J.J. (1975) The Probability Plot Correlation Coefficient Test for Normality. Technometrics. 17(1): 111-117.

Finney, D.J. (1971a) Statistical Method in Biological Assay. 2nd Ed. Griffin & Co. London.

Finney, D.J. (1971b) Probit Analysis. 3rd ed. Cambridge Univ. Press. Cambridge.

Fischhoff, B. (1982) For Those Condemned to Study the Past: Heuristics and Biases in Hindsight. IN Judgment Under Uncertainty: Heuristics and Biases. 335-351. D. Kahneman, P. Slovic and A. Tversky, eds. Cambridge Univ. Press.

Fisher, R.A. (1963) Statistical Methods for Research Workers. 13th ed. Oliver & Boyd, Pub. London.

Fisher, R.A. (1966) The Design of Experiments. Hafner Publ. NY.

Fleck, L. (1935/1979) Genesis and Development of a Scientific Fact. Univ. of Chicago Press. Chicago, IL.

Fleiss, J.L. (1973) Statistical Methods for Rates and Proportions. John Wiley & Sons.

Fowler, J.F. (1982) A Critical Look at Empirical Formulae in Fractionated Radiotherapy. IN: Biological Bases and Clinical Implications of Tumor Radioresistance. 201-204. G. Fletcher, et al, eds. Masson Pub. NY.

Fowler, J.F. (1983) Dose Response Curves for Organ Function or Cell Survival. Brit. J. Radiol. 56: 497-500.

Fowler, J.F. (1984a) What Next in Fractionated Radiotherapy? Br. J. Cancer. 49: Suppl. VI. 285-300.

Fowler, J.F. (1984b) Fractionated Radiation Therapy After Strandqvist. Acta Radiologica. 23: Fasc. 4. 209-216.

Fowler, J.F. (1989) The Linear-Quadratic Formula and Progress in Fractionated Radiotherapy. Br. J. Radiol. 62: 679-694.

Fowler, J.F. (1991) Carcinoma of the Lung: Hyperfractionation or Resection and Chemotherapy?

Int. J. Radia. Oncol. Biol. Phys. 20: 169-171.

Fowler, J.F. (1991) Apparent Rates of Proliferation of Acutely Responding Normal Tissues During Radiotherapy of Head and Neck Cancer. Int. J. Radia. Oncol. Biol. Phys. 21: 1451-1456.

Freeman, M.F. and Tukey, J.W. (1950) Transformations Related to the Angular and the Square Root. Annals of Math. Statis. 21: 607-611.

Friedman, P.J. (1988) Research Ethics, Due Process, and Common Sense. JAMA. 260(13): 1937-1938.

Frome, E.L. (1983) The Analysis of Rates Using Poisson Regression Models. Biometrics. 39(3): 665-674

Frome, E.L. (1986) Regression Methods for Binomial and Poisson Distributed Data. IN Multiple Regression Analysis: Applications in the Health Sciences. 84-123. D.E. Herbert and R.H. Myers, eds. AAPM Monograph No. 12. AIP. NY.

Frome, E.L. (In Press) Statistical Analysis of Cytogenetic Dose-Response Curves. To Appear in Statistical Methods in Toxicological Research. D. Krewski and C. Franklin, eds. Gordon & Breach Science Pubs. Inc.

Frome, E.L. and Beauchamp, J.J. (1968) Maximum Likelihood Estimation of Survival Curve Parameters. Biometrics. 595-605.

Frome, E.L. and Checkoway, H. (1985) Epidemiologic Programs for Computers and Calculators. Amer. J. Epidem. 121(2): 309-323.

Frome, E.L. and DuFrain, R.J. (1986) Maximum Likelihood Estimation for Cytogenetic Dose-Response Curves. Biometrics. 42: 73-84.

Frome, E.L., Kutner, M.H. and Beauchamp, J.J. (1973) Regression Analysis of Poisson-Distributed Data. Amer. Statis. Assn. 68: 935-940.

Fry, R.J.M. (1981) Experimental Radiation Carcinogenesis: What Have We Learned? Radia. Res. 87: 224-239.

Gaylord, D.W. and Merrill, J.A. (1968) Augmenting Existing Data in Multiple Regression. Technometrics. 10: 73-81.

Gehan, E.A. (1983) Comment on the "Ethical Guidelines for Statistical Practice: Report of the Ad Hoc Committee on Professional Ethics." Amer. Statistician. 37(1): 8-9.

Gehring, P.J., Watanabe, P.G. and Young, J.D. (1977) The Relevance of Dose-Dependent Pharmacokinetics in the Assessment of Carcinogenic Hazard of Chemicals. IN Origins of Human Cancer. Book A. Vol. 4: H.H. Hiatt, J.D. Watson and J.A. Winsten, eds. Cold Spring Harbor Lab.

Gilchrist, W. (1984) Statistical Modelling. John Wiley & Sons. NY.

Gilmore, R. (1992) Catastrophe Theory. Encyclopedia of Applied Phys. 3: 85-119. CH Pub. Inc.

Glantz, S.A. (1980) Biostatistics: How to Detect, Correct,and Prevent Errors in the Medical Literature. Circulation. 61: 1-7.

Glass, G.V. (1976) Primary, Secondary, and Meta-Analysis of Research. Ed. Research. 5(1): 3-8.

Glass, G., McGaw, B. and Smith, M.L. (1981) Meta-Analysis in Social Research. Sage Pub. Beverly Hills.

Gnanadesikan, R. (1977) Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons. NY.

Good, I.J. (1965) The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press. Research Monograph No. 30. Cambaridge, MA.

Graham, M.H. (1986) Expert Witness Testimony and the Federal Rules of Evidence: Insuring Adequate Assurance of Trustworthiness. Univ. of IL. Law Rev. 43: 43-90.

Greenwald, A.G. (1975) Consequences of Prejudice Against the Null Hypothesis. Psy. Bulletin. 82(1): 1-20.

Griffiths, W.E., Hill, R.C. and Pope, P.J. (1987) Small Sample Properties of Probit Model Estimators. J. Amer. Statis. Assn. 82: 929-937.

Hacking, I. (1983) Representing and Intervening. Cambridge Univ. Press. Cambridge.

Hacking, I. (1990) The Taming of Chance. Cambridge Univ. Press.

Hald, A. (1952) Statistical Theory with Engineering Applications. John Wiley & Sons.

Hall, E.J. (1975) Biological Problems in the Measurement of Survival at Low Doses. IN Cell Survival After Low Doses of Radiation: Theoretical and Clinical Implications. 13-24. Proc. 6th L.H. Gray Conf. London. Sept. 1974. Wiley & Sons. NY.

Hall, E.J. and Fowler, J.F. (1988) Radiobiology. Int. J. Radia. Oncol. Biol. Phys. 14: 525-528.

Halperin, M., Blackwelder, C. and Verter, J.I. (1971) Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. J. Chron. Dis. 24: 125-158.

Halvorsen, K.T. (1986) Combining Results from Independent Investigations. IN Medical Uses of Statistics. 392-416. J.C. Bailar and F. Mosteller, eds. NEJM. Waltham, MA.

Hamilton, D. (1990) Publishing by - and for? - the Numbers. Science. 250: 1331-1332.

Harris, R.J. (1975) A Primer of Multivariate Statistics. Academic Press. NY.

Hastings, N.A.J. and Peacock, J.B. (1975) Statistical Distributions. John Wiley and Sons. NY.

Hedges, L.V. and Olkin, I. (1985) Statistical Methods for Meta-Analysis. Academic Press. NY.

Hedrick, T.E. (1985) Justifications for and Obstacles to Data Sharing. IN Sharing Research Data. 123-148. S.E. Fienberg, M.E. Martin and M.L. Straf, eds. National Academy Press. Washington, DC.

Hensher, D.A. and Johnson, L.W. (1981) Applied Discrete-Choice Modelling. Halsted Press. London.

Herbert, D.E. (1978) Response Surfaces for Binary Events. An Empirical Basis for Radiation Oncology. IN Computers in Radiation Therapy. 210-228. Proceedings of the 6th Intl. Conf. on the Use of Computers in Radiation Therapy. Gottingen, Fed. Rep. of Germany. Sept. 18-23, 1977. U. Rosenow, ed. Mylet-Druck. Dransfeld.

Herbert, D.E. (1981) Battered Data - Some Clinical Effects of the Abuse of Multiple Regression Methods: The NSD. Med. Phys. 8(6): 813-847.

Herbert, D.E. (1983a) An Extreme Value Paradigm for the Effect of Size of Target Volume on End Results in Radiation Oncology. Med. Phys. 10(5): 589-604.

Herbert, D.E. (1983b) Model or Metaphor? More Comments on the BEIR III Report. Epidemiology Applied to Health Physics. 357-390. Proceedings of the Health Physics Society 16th Midyear Topical Meeting. Albuquerque, NY.

Herbert, D.E. (1985a) A General Linear Model of Clinical Radiation Effects. IN Optimization of Cancer Radiotherapy. 226-296. B.R. Paliwal, D.E. Herbert and C.G. Orton, eds. AAPM Symposium Proceedings No. 5. AIP. NY.

Herbert, D.E. (1985b) Two Commentaries on Modelling of Clinical Response. A Model of "Modelling" and the Law of Small Numbers. IN Optimization of Cancer Radiotherapy. 527-538. B.R. Paliwal, D.E. Herbert and C.G. Orton, eds. AAPM Symposium No. 5. AIP. NY.

Herbert, D.E. (1985c) An Empirical, "Extreme Value" Model of the Volume Effect in Radiation Oncology. IN Optimization of Cancer Radiotherapy. 381-401. B.R. Paliwal, D.E. Herbert and C.G. Orton, eds. AAPM Symposium Proc. No. 5. AIP. NY.

Herbert, D.E. (1986a) Clinical Dose Response Models. I. Regression Diagnostics and Biased Estimation. IN Multiple Regression Analysis: Applications in the Health Sciences. 208-306. D.E. Herbert and R.E. Myers, eds. AAPM Monograph No. 13. AIP. NY.

Herbert, D.E. (1986b) Clinical Dose-Response Models II. Probit and Logit Models of Experimental and Non-Experimental Data. IN Ibid. 385-453.

Herbert, D.E. (1986c) Clinical Radiocarcinogenesis. Applications of Regression Diagnostics and Bayesian Methods to Poisson Regression Models. IN Ibid. 307-364.

Herbert, D.E. (1986d) Clinical Diagnostic Models. Discriminant Analysis In Vitro NMR Measurements on Normal and Malignant Tissues. IN Ibid. 487-545.

Herbert, D.E. (1986e) The Perceived Clinical Value of NMR Measurements on Biopsy Specimens. Part I. Interval Estimates of Diagnostic Error Rates and a Note on an Effect of the Law of Small Numbers. Magnetic Resonance Imaging. 4: 215-227.

Herbert, D.E. (1987) Critical Review. "What next in Fractionated Radiotherapy?" by J. F. Fowler. Invest. Radiol. 22(5): 447-450.

Herbert, D.E. (1989a) Some Impressions of the 3rd ICDTF. A Newer Organon? IN: Prediction of Response in Radiation Therapy: The Physical and Biological Basis. (Part 1) 367-377. B. Paliwal, J. Fowler, D. Herbert, T. Kinsella, and C. Orton, eds. AAPM Sympo. Proc. No. 7. AIP. NY.

Herbert, D.E.(1989b) Reflections on the LQ Model. Does it "Fit" (Does it Matter?). IN: Prediction of Response in Radiation Therapy: Analytical Models and Modelling. (Part 2) Ibid. 400-516.

Herbert, D.E. (1989c) Dose-Response Models: Construction, Criticism, Discrimination, Validation, and Deployment. Ibid. 534-630.

Herbert, D.E. (1989d) Some Applications of Semi-Bayesian Methods to Dose-Response Modelling. Ibid. 660-717.

Herbert, D.E. (1993a) Time Factors in Human Tumors: What Do We Know - And When Did We Know It? How We Can Learn From Data (Or Fail To). IN Prediction of Response in Radiation Therapy: Radioresistance and Repopulation. (In press) Proceedings of the 4th Intl. Conf. on Dose, Time, and Fractionation. Madison, WI. Sept. 1992.

Herbert, D.E. (1993b) Overview of Some Useful Concepts, Methods, and Criteria New to Radiobiological Modelling: 1) Bivariate Probit Models of Probability of Uncomplicated Local Control of Tumor. 2) An Example of the Use of the Stein Effect. In Ibid.

Hewlett, P.S. and Plackett, R.L. (1979) The Interpretation of Quantal Responses in Biology. Edward Arnold Pub. London.

Hibben, J.G. (1984) Hegel's Logic. Garland Pub. NY.

Higman, B. (1955) Applied Group-Theoretic and Matrix Methods. Clarendon Press. Oxford.

Hillcoat, B.L. (1984) Data Recycling and Misreading: Two Potential Errors in Pooled Data from Small Studies. J. Clinical Oncol. 2(9); 1047-1049.

Hinde, J. (1982) Compound Poisson Regression Models. IN GLIM 82: Proceedings of the International Conference on Generalized Linear Models. 109-121. R. Gilchrist, ed. Springer-Verlag. NY.

Hinkley, D.V. (1977) Jackknifing in Unbalanced Situations. Technometrics. 19(3): 285-292.

Hoadley, A.B. and Kettenring, J.R. (1990) Communications Between Statisticians and Engineers/ Physical Scientists. Technometrics. 32(3): 243-274.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1985) Exploring Data Tables, Trends, and Shapes. John Wiley and Sons. NY.

Hocking, R.R. (1983) Developments in Linear Regression Methodology: 1959-1982 (with discussion). Technometrics. 25(3): 219-249.

Hodges, S.D. and Moore, P.G. (1972) Data Uncertainties and Least Squares Regression. Applied Statis. 21(2): 185-195.

Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Applications to Non-orthogonal Problems. Technometrics. 12: 69-82.

Holton, G. (1978) The Scientific Imagination: Case Studies. Cambridge Univ. Press.

Hoos, I. (1980) Risk Assessment in Social Perspective. IN Perceptions of Risk. 57-84. NCRP Proceedings #1. Proceedings of the Fifteenth Annual Meeting of the NCRP. Washington, DC.

Horrobin, D.F. (1990) The Philosophical Basis of Peer Review and the Supression of Innovation. JAMA. 263: 1438-1441.

Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. John Wiley & Sons. NY.

Hosmer, D.W., Lemeshow, S. and Klar, J. (1988) Goodness-of-fit Testing for the Logistic Regression Model when the Estimated Probabilities are Small. Biometrical J. 30: 911-924.

Huber, P.J. (1977) Robust Statistical Procedures. SIAM. Philadelphia.

Huber, P.J. (1981) Robust Statistics. John Wiley & Sons. NY.

Huck, S.W. and Sandler, H.M. (1979) Rival Hypotheses. Harper & Row. NY.

Hume, D. (1745/1977) An Enquiry Concerning Human Understanding: A Letter from a Gentleman to his Friend in Edinburgh. E. Steinberg, ed. Hackett Pub. Indianapolis,IN.

Hunter, W.G. and Reiner, A.M. (1965) Designs for Discriminating Between two Rival Models. Technometrics. 7(3): 307-323.

Jacobsen, M. (1976) Against Popperized Epidemiology. Intl. J. Epidem. 5(1): 9-11.

Jaffe, A.J. and Spirer, H.F. (1987) Misused Statistics. Dekker Pub. NY.

Jeffreys, H. (1957) Scientific Inference. 2nd ed. Cambridge Univ. Press.

Jeffreys, H. (1961) Theory of Probability. 3rd ed. Clarendon Press. Oxford.

Johnson, N.L. and Leone, F.C. (1964) Statistics and Experimental Design. 88-91. Vol. 1: Wiley & Sons. NY.

Johnston, J. (1972) Econometric Methods. 2nd ed. McGraw-Hill Book Co. NY.

Jones, G. and Laukkanen, E. (1987) Tolerance Revisited. Intl. J. Radia. Oncol. Biol. Phys. 13: 290-291.

Kahn, H.A. and Sempos, C.T. (1989) Statistical Methods in Epidemiology. Oxford Univ. Press. NY.

Kahneman, D., Slovic, P. and Tversky, A., eds. (1982) Judgment Under Uncertainty: Heuristics and Biases. Cambridge Univ. Press. Cambridge.

Kalbfleisch, J.D. and Prentice, R.L. (1980) The Statistical Analysis of Failure Time Data. John Wiley & Sons. NY.

Kant, I. (1800) Logic.

Keane, T.J., Fyles, A., O'Sullivan, B., Barton, M., Maki, E., and Simm, J. (1992) The Effect of Treatment Duration on Local Control of Squamous Carcinoma of the Tonsil and Carcinoma of the Cervix. Seminars of Radia. Oncol. 2: 26-28.

Kellert, S.H. (1993) In the Wake of Chaos. Univ. of Chicago Press.

Kendall, M.G. (1966) Discrimination and Classification. IN Multivariate Analysis. 165-186. P. Krishnaiah, ed. Academic Press. NY.

Kendall, M.G. and Buckland, W.R. (1971) A Dictionary of Statistical Terms. Hafner Pub. Co. NY.

Kennard, R.W. and Stone, L.W. (1969) Computer Aided Design of Experiments. Technometrics. 11: 137-148.

Kennedy, A.R. (1985) Relevance of Tumor Promotion to Carcinogenesis in Human Populations. IN Carcinogenesis. Vol. 8. 431-435. M.J. Mass, et al. Raven Press. NY.

Kerkut, G.A. (1983) Choosing a Title for a Paper. Comp. Biochem. Physiol. 74A: 1.

Klein, S., Simes, J. and Blackburn, G.L. (1986) Total Parenteral Nutrition and Cancer Clinical Trials. Cancer. 58(6): 1378-1386.

Koestler, A. (1964) Act of Creation. MacMillan. NY.

Kolakowski, D. and Bock, R.D. (1981) A Multivariate Generalization of Probit Analysis. Biometrics. 37: 541-551.

Kotz, S. and Johnson, N.L., eds. (1982) Encyclopedia of Statistical Sciences. John Wiley & Sons. NY.

Kuhn, T. (1970a) The Structure of Scientific Revolutions. 2nd ed. Univ. of Chicago Press. Chicago.

Kuhn, T. (1970b) Logic of Discovery or Psychology of Research? IN Criticism and the Growth of Knowledge. I. Lakatos & A. Musgrave, eds. 1-25. Cambridge Univ. Press. Cambridge.

Kuhn, T. (1970c) Reflections on my Critics. In Criticism and the Growth of Knowledge. 231-278. I. Lakatos & A. Musgrave, eds. Cambridge Univ. Press.

Kuhn, T. (1977) The Essential Tension. Univ. of Chicago Press. Chicago.

Kyburg, H. (1970) Comments on "Occam's Razor Needs New Blades - by H. Rubin. IN: Foundations of Statistical Inference. 375-376. V.P. Godambe and D.A. Sprott, eds. Sprott, Holt, Rinehart, and Winston of Canada, Ltd. Toronto.

L'Abbe, K.A., Detsky, A.S., and O'Rourke, K. (1987) Meta-Analysis in Clinical Research. Annals Internal Med. 107: 224-233.

Lachenbruch, P.A. (1975) Discriminant Analysis. Hafner Press. N.Y.

Land, C.E. (1981) Biological Models in Epidemiology: Radiation Carcinogenesis. IN Environmental Science Research. 21: 71-104. G. Berg and H.E. Maille. NY.

Land, C.E., Boice, J.D., Shore, R.E., Norman, J.E. and Tokunaga, M. (1980) Breast Cancer Risk from Low-Dose Exposures to Ionizing Radiation: Results of Parallel Analysis of Three Exposed Populations of Women. JNCI 65(2): 353-376.

Landi, G. and Ciccone, A. (1993) Publication Bias Via Suppressed Criticism. Lancet. 341: 697-698.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984) Graphical Methods for Assessing Logistic Regression Models (with comments). J. Amer. Statis. Assn. 79: 61-83.

Lawless, J.F. (1982) Statistical Models and Methods for Lifetime Data. John Wiley & Sons. NY.

Lawless, J.F. (1987) Regression Methods for Poisson Process Data. J. Amer. Statis. Assn. 82: 808-815.

Leamer, E.E. (1978) Specification Searches. John Wiley & Sons. NY.

Leamer, E.E. (1982) Sets of Posterior Means with Bounded Variance Priors. Econometrics. 50(3): 725-736.

Leamer, E.E. (1983) Let's Take the Con Out of Econometrics. Amer. Economic Review. 73: 31-43.

Leamer, E.E. (1986) Bayesian Regression and Sensitivity Analyses. IN Multiple Regression Analysis: Applications in the Health Sciences. 58-74. D.E. Herbert and R.E. Myers, eds. AAPM Monograph No. 13. AIP. NY.

Lemeshow, S. and Hosmer, D.W. (1982) A Review of Goodness-of-fit Statistics for Use in Development of Logistic Regression Models. Amer. J. Epidem. 115: 92-106.

Lemmer, H.H. (1981) From Ordinary to Bayesian Shrinkage Estimators. South African Statis. J. 15: 57-72.

Lesaffre, E. and Molenberghs, G. (1991) Multivariate Probit Analysis: A Neglected Procedure in Medical Statistics. Statis. in Med. 10: 1391-1403.

Lindeman, R.H., Merenda, P. and Gold, R. (1980) Introduction to Bivariate and Multivariate Analysis. Scott, Foreman and Co. Glenview, IL.

Lindley, D.V. (1968) The Choice of Variables in Multiple Regression. J. Royal Statis. Soc. 30: 31-66.

Lindley, D.V. and Beale, E.M.L. (1968) Discussion of Lindley's paper, The Choice of Variables in Multiple Regression. J. Royal Statis. Soc. 30: 54-56.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes Estimates for the Linear Model. J. Royal Statis. Soc. B. 34(1): 1-41.

Linhart, H. and Zucchini, W. (1986) Model Selection. John Wiley & Sons. NY.

Louis, T.A., Fineberg, H.V. and Mosteller, F. (1985) Findings for Public Health from Meta-Analyses. Ann. Rev. Public Health. 6: 1-20.

Lubin, A. (1950) Linear and Non-Linear Discriminating Functions. Br. J. Psychology. Statis. Sec. 3: Part 2. 90-104.

Mah, K., Van Dyk, J., Keane, T. and Poon, P. (1987) Acute Radiation-Induced Pulmonary Damage: A Clinical Study on the Response to Fractionated Radiation Therapy. Int. J. Radia. Oncol. Biol. Phys. 13: 179-188.

Mahoney, M.J. (1977) Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. Cognitive Therapy and Research. 1(2): 161-175.

Maindonald, J.H. (1984) Statistical Computation. John Wiley & Sons. NY.

Malaise, E.P., Fertil, B., Deschavanne, P.J., Chavaudra, N. and Brock, W.A. (1987) Initial Slope of Radiation Survival Curves is Characteristic of the Origin of Primary and Established Cultures of Human Tumor Cells and Fibroblasts. Radia. Res. 111: 319-333.

Malinvaud, E. (1980) Statistical Methods of Econometrics. 3rd ed. North-Holland. NY.

Margolin, B.H., Kaplan, N. and Zeiger, E. (1981) Statistical Analysis of the Ames Salmonella/Microsome Test. Proc. Natl. Acad. Sci. USA. 78(6); 3779-3783.

Marks, R., Dawson-Saunders, E.K., Bailar, J.C., Dan, B.B. and Verran, J.A. (1988) Interactions Between Statisticians and Biomedical Journal Editors. Statis. in Med. 7: 1003-1011.

Marquardt, D.W. (1980) Comment on G. Smith & F. Campbell article, "A Critique of Some Ridge Regression Methods". J. Amer. Statis. Assn. 75: 87-91.

Marquardt, D.W. (1987) The Importance of Statisticians. J. Amer. Statis. Assn. 82: 1-7.

Marquardt, D.W. and Snee, R.D. (1975) Ridge Regression in Practice. Amer. Statis. 29(1): 3-19.

Marshall, J.R. (1990) Data Dredging and Noteworthiness. Epidem. 1: 5-7. 1990.

Maslow, A. (1966) The Psychology of Science: A Reconnaissance. Harper & Row. NY.

Mazur, A. (1973) Disputes Between Experts. Minerva: A Review of Science, Learning and Policy. 11: 243-262.

McCade, G.P. (1978) Evaluation of Regression Coefficient Estimates Using-Acceptability. Technometrics. 20(2): 131-139.

McCormmach, R., ed. (1974) Historical Studies in the Physical Sciences. 4th Annual Vol. Princeton Univ. Press. Princeton. NJ.

McCullagh, P. and Nelder, J.A. (1983) Generalized Linear Models. Chapman & Hall. NY.

McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd ed. Chapman & Hall, NY.

McDonald, G.C. and Galarneau, D.I. (1975) A Monte Carlo Evaluation of Some Ridge-Type Estimators. J. Amer. Statis. Assn. 70: 407-416.

McNeil, K.A. and Spaner, S.D. (1971) Brief Report: Highly Correlated Predictor Variables in Multiple Regression Models. Multivariate Behavioral Res. 6: 117-125.

Medicare Hospital Information Report (Alabama) (1992) Vol. 1. U.S. Govern. Printing Off. Superintendent of Documents. Washington, DC.

Metz, C.E., Tokars, R. P., Kronman, K.B.and Griem, M.L. (1982) Maximum Likelihood Estimation of Dose-Response Parameters for Therapeutic Operating Characteristic (TOC) Analysis of Carcinoma of the Nasopharynx. Int. J. Radia. Oncol. Biol. Phys. 8: 1185-1192.

Millar, B.C., Fielden, E.M. and Millar, J.L. (1978) Interpretation of Survival-Curve Data for Chinese Hamster Cells, line V-79 using the Multi-Target, Multi-Target with Initial Slope, and α, β equations. Intl. J. Radia. Biol. 33(6): 599-603.

Miller, D.M. (1984) Reducing Transformation Bias in Curve Fitting. Amer. Statis. 38: 124-126.

Miller, R.G. (1974) An Unbalanced Jackknife. Annals of Statis. 2(5): 880-891.

Minkin, S. (1987) Optimal Designs for Binary Data. J. Amer. Statis. Assn. 82: 1098-1103.

Montgomery, D.C. and Peck, E.A. (1982) Introduction to Linear Regression Analysis. John Wiley & Sons. NY.

Montour, J.L., Hard, R.C. and Flora, R.E. (1977) Mammary Neoplasia in the Rat Following High-Energy Neutron Irradiation. Cancer Research. 37: 2619-2623.

Moon, F.C. (1992) Chaotic and Fractal Dynamics. John Wiley & Sons. NY.

Moore, D.H. and Mendelsohn, M.L. (1972) Optimal Treatment Levels in Cancer Therapy. Cancer. 30: 97-106.

Mosteller, F., Siegel, A.F., Trapido, E. and Youtz, C. (1981) Eye Fitting Straight Lines. Amer. Statis. 35(3): 150-152.

Mosteller, F. and Tukey, J.W. (1977) Data Analysis and Regression. Addison-Wesley Pub. Co. Reading, MA.

Muirhead, C.R. and Darby, S.C. (1987) Modelling the Relative and Absolute Risks of Radiation-Induced Cancers. J. Royal Statis. Soc. A: 150: Part 2. 83.

Murphy, E.A. (1976) The Logic of Medicine. John Hopkins Univ. Press. Baltimore, MD.

Murphy, E.A. (1982) The Analysis and Interpretation of Experiments: Some Philosophical Issues. J. Med. and Philos. 7: 307-325.

Myers, M.H., Axtell, L.M. and Zelen, M. (1966) The use of Prognostic Factors in Predicting Survival for Breast Cancer Patients. J. Chron. Dis. 19: 923-933.

Myers, R.H. (1971) Response Surface Methodology. Allyn and Bacon, Inc. Boston, MA.

Myers, R.H. (1990) Classical and Modern Regression with Applications. PWS-Kent Pub. Co. Boston, MA.

Narin, F. and Frame, J. (1989) The Growth of Japanese Science and Technology. Science. 245: 600-605.

NAS/NRC (1980) The Effects on Populations of Exposure to Low Levels of Ionizing Radiation: 1980. (BEIR III) National Academy Press. Washington, DC.

NAS/NRC (1988) Health Risks of Radon and Other Internally Deposited Alpha-Emitters. (BEIR

IV) National Academy Press. Washington, DC.

NAS/NRC (1990) Health Effects of Exposure to Low Levels of Ionizing Radiation. (BEIR V) National Academy Press. Washington, DC.

NAS/NRC (1992) Combining Information. Statistical Issues and Opportunities for Research. Washington, DC.

NCRP (1980) Influence of Dose and Its Distribution in Time on Dose-Response Relationships for Low-Let Radiations. NCRP Report No. 64. Washington, DC.

Nelder, J.A. (1968) Regression, Model-building and Invariance. J. Royal Statis. Soc. Series A. 131: 303-329.

Nelder, J.A. (1986) Statistics, Science and Technology. J. Royal Statis. Soc. A. 149: part 2, 109-121.

Nelder, J.A. (1990) Nearly Parallel Lines in Residual Plots. Amer. Statis. 44(3): 221-222.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Liner Models. J. Royal Statis. Soc. A. 135: Part 3. 370-384.

Neter, J. and Wasserman, W. (1974) Applied Linear Statistical Models. Richard H. Irvin Pub. Homewood, IL.

Neufeld, A.H. (1986) Reproducing Results. Science. 234: 11.

Neugebauer, E., Lorenz, W., Maroske, D., Barthlen, W. and Ennis, M. (1987) The Role of Mediators in Septic/Endotoxic Shock. Theor. Surgery. 2: 1-28.

NIH (1985) Report of the National Institutes of Health Ad Hoc Working Group to Develop Radioepidemiological Tables. NIH Pub. No. 85-2748. U.S. Govt. Printing Office. Washington, DC.

Nicolis, G. and Prigogine, I. (1989) Exploring Complexity. W.H. Freeman, Pub.

Niemierko, A., Urie, M. and Gottein, M. (1992) Optimization of 3D Radiation Therapy with Both Physical and Biological End Points and Constraints. Int. J. Radia. Oncol. Biol. Phys. 23: 99-108.

Nisbett, R. and Ross, L. (1980) Human Inference: Strategies and Shortcomings of Social Judgment. Prentice-Hall Inc. Englewood Cliffs, NJ.

Oakes, M. (1990) Statistical Inference. Epidem. Resources, Inc., Chestnut Hill, MA.

Obenchain, R.L. (1977) Classical F-Tests and Confidence Regions for Ridge Regression. Technometrics. 19(4): 429-439.

Obenchain, R.L. (1980) Comment on article, "A Critique of Some Ridge Regression Methods" by G. Smith & F. Campbell. J. Amer. Statis. Assn. 75: 95-96.

Oman, S.D. (1978) A Bayesian Comparison of Some Estimators Used in Linear Regression with Multicollinear Data. Commun. Statis. - Theor. Meth. A7. 6: 517-534.

Overgaard, M. (1988) Spontaneous Radiation-Induced Rib Fractures in Breast Cancer Patients Treated with Postmastectomy Irradiation. Acta Oncol. 27: Fasc. 2. 117-122.

Park, C.N. and Snee, R.D. (1983) Quantitative Risk Assessment: State-of-the-Art for Carcinogenesis. Amer. Statis. 37(4): 427-441.

Pearson, K. (1892/1957) The Grammar of Science. Meridian Books, Inc. NY.

Peirce, C.S. (1955) Philosophical Writings of Peirce. ed. by J. Buchler. Dover Pub. NY.

Pennsylvania Health Care Cost Containment Council (1992) Coronary Artery Bypass Graft Surgery. A Consumer Guide To. Harrisburg Transportation Center. Harrisburg, PA.

Peto, R. (1977) Epidemiology, Multistate Models, and Short-term Mutagenicity Tests. IN Origins of Human Cancer. 1403-1428. H.H. Hiatt, J.D. Watson and J. A. Winsten, eds. Cold Spring Harbor Laboratory. NY

Peto, R. (1979) Detection of Risk of Cancer to Man. Proc. Royal Society London. 205: 111-120.

Pippard, A.B. (1985) Response and Stability. Cambridge Univ. Press. Cambridge.

Pitman, E.J.G. (1979) Some Basic Theory for Statistical Inference. Chapman and Hall. London.

Platt, J.R. (1964) Strong Inference. Science. 146: 347-353.

Pocock, S.J. and Hughes, M.D. (1990) Estimation Issues in Clinical Trials and Overviews. Statis. in Med. 9: 657-671.

Popper, K. (1965a) The Logic of Scientific Discovery. Harper and Row Pub. Inc. NY.

Popper, K. (1965b) Conjectures and Refutations: The Growth of Scientific Knowledge. Harper & Row. NY.

Poston, T. (1979) The Elements of Catastrophe Theory or the Honing of Occam's Razor. IN: Transformations: Mathematical Approach to Culture Change. 425-436. C. Renfrew and K. Cooke, eds. Academic Press.

Pregibon, D. (1981) Logistic Regression Diagnostics. Annals of Statis. 9(4): 705-724.

Press, S.J. and Wilson, S. (1978) Choosing Between Logistic Regression and Discriminant Analysis. J. Amer. Statis. Assn. 73: 699-705.

Preston, D. (1989) Modeling Radiation Effects on Disease Incidence. Presented at the ASA Conf. on Radiation and Health, VIII. July 9-13, 1989. Copper Mountain, CO.

Price, D.J.D. (1963) Little Science, Big Science. Columbia Univ. Press. NY.

Prigogine, I. (1980) From Being to Becoming. Freeman & Co. NY.

Purchase, I.F.H. (1980) Inter-Species Comparisons of Carcinogenicity. Br. J. Cancer. 41:

Raiffa, H. (1982) Science and Policy: Their Separation and Integration in Risk Analysis. - Amer. Statis. 36(3): Part 2. 225-231.

Rao, C.R. (1973) Linear Statistical Inference and Its Applications. 2nd ed. Wiley & Sons. NY.

Ratkowsky, D.A. (1983) Nonlinear Regression Modelling. Marcel Dekker, Inc. NY.

Ratkowsky, D.A. (1990) Handbook of Nonlinear Regression Models. Marcel Dekker, Inc. NY.

Raudenbush, S.W. and Bryk, A.S. (1985) Empirical Bayes Meta-Analysis. J. Ed. Statis. 10(2): 75-98.

Relman, A. (1988) Assessment and Accountability. The Third Revolution in Medical Care. NEJM. 319: 1220-1222.

Robins, J.M. and Greenland, S. (1986) The Role of Model Selection in Causal Inference from Nonexperimental Data. Amer. J. Epidem. 123(3): 392-402.

Rolph, J.E. (1976) Choosing Shrinkage Estimators for Regression Problems. Commun. Statis. A5. 789-802.

Roper, W.L., Winkenwerder, W., Hackbarth, G.M. and Krakauer, H. Effectiveness in Health Care. NEJM. 319. 1197-1202.

Rosen, J. (1983) A Symmetry Primer for Scientists. John Wiley & Sons. NY.

Rosenfeld, A.H. (1975) The Particle Data Group: Growth and Operations - Eighteen Years of Particle Physics. Annual Rev. of Nucl. Sci. 25: 555-595.

Rosenthal, R. (1979) The File Drawer Problem and Tolerance for Null Results. Psy. Bulletin. 86: 638-641.

Ross, L. and Lepper, M.R. (1980) The Perseverance of Beliefs: Empirical and Normative Considerations. New Directions for Method. of Social and Behav. Sci. 4: 17-36.

Rothman, K.J. (1978) Occam's Razor Pares the Choice Among Statistical Models. Amer. J. Epidem. 108(5): 347-349.

Rousseeuw, P.J. and van Zomeren, B.C. (1990) Unmasking Multivariate Outliers and Leverage Points (with Discussions). J. Amer. Stat. Assn. 85: 633-651.

Rubin, P. and Casarett, G. (1972) A Direction for Clinical Radiation Pathology: The Tolerance Dose. Frontiers Radia. Thera. Oncol. 6: 1-16.

Ruelle, D. (1991) Chance and Chaos. Princeton Univ. Press.

Sacks, H.S., Berrier, J., Rettman, D., Ancona-Bert, V.A. and Chalmers, T.C. Meta-Analyses of Randomized Controlled Trials. NEJM. 316: 450-455.

Salmon, W.C. (1966) The Foundations of Scientific Inference. Univ. of Pittsburgh Press, PA.

Salsburg, D. (1971) Testing Dose Responses on Proportions Near Zero or One with the Jackknife. Biometrics. 27: 1035-1041.

Schaefer, R.L. (1984) Alternative Estimators in Logistic Regression When the Data Are Collinear. J. Computation and Simulation.

Schaefer, R.L., Roi, L.D. and Wolfe, R.A. (1984) A Ridge Logistic Estimator. Commun. Statist. - Theor. Meth. 13(1): 99-113.

Schaffner, K.F. (1985) Logic of Discovery and Diagnosis in Medicine. Univ. Calif. Press. Berkeley, CA.

Schmaus, W. (1988) An Analysis of Fraud and Misconduct in Science. IN Project on Scientific Fraud and Misconduct. 87-115. AAAS-ABA Natl. Conf. of Lawyers and Scientists. Washington, DC.

Schor, S. and Karten, I. (1966) Statisstical Evaluation of Medical Journal Manuscripts. JAMA. 195: 145-150.

Schultheiss, T.E. (1981) Comments on Optimization of Radiotherapy. Amer. J. Clin. Oncol. Cancer Clin. Trials. 4: 559.

Schultheiss, T.E. (1987) Tolerance: A Useful Concept in Radiation Therapy. Int. J. Radia. Oncol. Biol. Phys. 13: 470.

Schultheiss, T.E. and Orton, C.G. (1985) Bioeffect Optimization: Decision Theory Model. Optimization of Cancer Radiotherapy. 471-478. Proc. of the 2nd Int'L Conf. on Dose, Time and Fractionation in Radiation Oncology. Madison, WI. Sept. 1984. B. Paliwal, D. Herbert, & C. Orton, (eds) AAPM Sympo. Proc. No. 5.

Schultheiss, T.E., Orton, C.G. and Peck, R.A. (1983) Models in Radiotherapy Volume Effects. Med. Phys. 10(4): 410-415.

Schwarz, G. (1978) Estimating the Dimension of a Model. Annals of Statis. 6(2): 461-464.

Searle, S.R. (1982) Matrix Algebra Useful for Statistics. John Wiley & Sons. NY.

Searle, S.R. (1988) Parallel Lines in Residual Plots. Amer. Statis. 42(3): 244.

Seber, G.A.F. (1977) Linear Regression Analysis. John Wiley & Sons. NY.

Shafer, G. (1976) A Mathematical Theory of Evidence. Princeton Univ. Press. N.J.

Shellabarger, C.J., Bond, V.P., Cronkite, E.P and Aponte, G.E. (1969) Relationship of Dose of Total-Body 60Co Radiation to Incidence of Mammary Neoplasia in Female Rats. IN Radiation Induced Cancer. IAEA. Vienna. 161-172.

Shellabarger, C.J., Stone, J.P. and Holtzman, S. (1986) Experimental Carcinogenesis in the Breast. IN Radiation Carcinogenesis. 169-180. A.C. Upton, R.E. Albert, F.J. Burns and R.E. Shore, eds. Elsevier. NY.

Simes, R.J. (1987) Confronting Publication Bias: A Cohort Design for Meta-Analysis. Statis. in Med. 6: 11-29.

Simpson, E.H. (1951) The Interpretation of Interaction in Contingency Tables. J. Royal Statis. Soc. B. 13(2): 238-241.

Smart, R.G. (1964) The Importance of Negative Results in Psychological Research. The Canadian Psychologist. 5(4): 225-232.

Smith, G. and Campbell, F. (1980) A Critique of Some Ridge Regression Methods (with comments). J. Amer. Statis. Assn. 75: 74-81.

Smith, M.L. (1980) Publication Bias and Meta-Analysis. Eval. in Ed. 4: 22-24.

Snee, R.D. (1973) Some Aspects of Nonorthogonal Data Analysis. Part I. Developing Predictive Equations. J. Qual. Tech. 5(2): 67-79.

Snee, R.D. (1977) Validation of Regression Models: Methods and Examples. Technometrics. 19(4): 415-428.

Snee, R.D. (1986) An Alternative Approach to Fitting Models When Re-Expression of the Response is Useful. J. Quality Tech. 18(4): 211-225.

Snee, R.D. and Irr, J.D. (1981) Design of a Statistical Method for the Analysis of Mutagenesis at the Hypoxanthine-Guanine Phosphoribosyl Transferase Locus of Cultured Chinese Hamster Ovary Cells. Mutation Res. 85: 77-93.

Snee, R.D. and Marquardt, D.W. (1984) Collinearity Diagnostics Depend on the Domain of Prediction, the Model, and the Data. The Amer. Statis. 38(2): 83-87.

Sommerfeld, A. (1949) Partial Differential Equations in Physics. Academic Press. NY.

Sparrow, A.H., Underbrink, A.G. and Rossi, H.H. (1972) Mutations Induced in Tradescantia by Small Doses of X-rays and Neutrons: Analysis of Dose-Response Curves. Science. 176: 916-918.

Stein, C. (1955) Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. Proc. of the 3rd Berkeley Sympo. on Math. Statis. and Prob. 1: 197-206.

Berkeley. Univ. of Calif. Press.

Steinberg, D. (1989) Induced Work Participation and the Returns to Experience for Welfare Women. J. Econometrics. 41: 321-340.

Steinberg, D.M. and Hunter, W.G. (1984) Experimental Design: Review and Comment. Technometrics. 26(2): 71-130.

Sterling, T.D. and Weinkam, J.J. (1979) What Happens when Major Errors are Discovered Long After an Important Report has been Published? Paper presented at the annual meeting of the American Statistical Association, August 1979.

Stewart, I. (1989) Does God Play Dice? Blackwell, Worcester, UK.

Stewart, W. and Feder, N. (1987) The Integrity of the Scientific Literature. Nature. 325: 207-214.

Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. J. Royal Statis. Soc. B. 36: 111-147.

Stone, M. (1976) An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. Science. 8(1): 44-47.

Supe, S.J., Nagalaxmi, K.V. and Meenaks, L. (1983) Tumor Significant Dose. Med. Phys. 10(1): 51-56.

Susser, M. (1986) The Logic of Sir Karl Popper and the Practice of Epidemilogy. Amer. J. Epidem. 146: 711-718.

Swindel, B.F. (1976) Good Ridge Estimators Based on Prior Information. Commun. Statist. - Theor. Meth. A5(11): 1065-1075.

Thames, H.D. (1988) Early Fractionation Methods and the Origins of the NSD Concept. Acta Oncol. 27: Fasc. 2. 89-103.

Thames, H.D., Withers, H.R., Peters, L.J. and Fletcher, G.H. (1982) Changes in Early and Late Radiation Responses with Altered Dose Fractionation: Implications for Dose-Survival Relationships. Int. J. Radia. Oncol. Biol. Phys. 8: 219-226. Theil, H. (1971) Principles of Econometrics. John Wiley & Sons. NY.

Theil, H. and Goldberger, A.S. (1961) On Pure and Mixed Statistical Estimation in Economics. Intl. Economic Rev. 2: 65-78.

Thisted, R.A. (1980) Comment on article, "A Critique of Some Ridge Regression Methods" by G. Smith & F. Campbell. J. Amer. Statis. Assn. 75: 81-86.

Thompson, J.M.T. and Stewart, H.B. (1986) Nonlinear dynamics and Chaos. John Wiley & Sons. NY.

Thorne, M.C. (1987) Principles of the International Commission on Radiological Protection System of Dose Limitation. Br. J. Radiol. 60: 32-38.

Till, J.E. and McCulloch, E.A. (1961) A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. Radia. Res. 14: 213-222.

Tong, H. (1977) Some Comments on the Canadian Lynx Data. J. Royal Statis. Soc. A. 140 Part 4. 432-436.

Toulmin, S.E. (1970) Does the Distinction Between Normal and Revolutionary Science Hold Water? Criticism and the Growth of Knowledge. 39-48. I. Lakatos and A. Musgrave, eds. Cambridge Univ. Press. Cambridge.

Touloukian, Y.S. (1975) Reference Data on Thermophysics. Thermochem. & Thermody. 10: 119-146.

Travis, E.L. and Tucker, S.L. (1987) Isoeffect Models and Fractionated Radiation Therapy. Int. J. Radia. Oncol. Biol. Phys. 13: 283-287.

Truett, J., Cornfield, J. and Kannel, W. (1967) A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. J. Chron. Dis. 20: 511-524.

Tucker, S.L. (1984) Tests for the Fit of the Linear-Quadratic Model to Radiation Isoeffect Data. Int. J. Radia. Oncol. Biol. Phys. 10: 1933-1939.

Tucker, S.L. and Thames, H.D. (1983) Flexure Dose: the Low-Dose limit of Effective Fractionation. Int. J. Radia. Oncol. Biol. Phys. 9: 1373-1383.

Tukey, J.W. (1969) Analyzing Data: Sancfification or detection work? Amer. Psychologist. 83-91.

Tukey, J.W. (1979) Methodology, and the Statistician's Responsibility for Both Accuracy and

Relevance. J. Amer. Statis. Assn. 74: 786-793.

Tukey, J.W. (1986) Sunset Salvo. Amer. Statistician. 40(1): 72-76.

Tversky, A. and Kahneman, D. (1982a) Belief in the Law of Small Numbers. IN Judgment Under Uncertainty: Heuristics and Biases. 23-31. D. Kahneman, P. Slovic and A. Tversky, eds. Cambridge Univ. Press. Cambridge.

Tversky, A. and Kahneman, D. (1982b) Causal Schemas in Judgements Under Uncertainty. Ibid. 117-128.

U.S. Congress, Office of Technology Assessment. (1988) The Quality of MedicalCare: Information for Consumers. OTA-H-386 (Washington, DC: U.S. Govern. Printing Off. June 1988).

Ullrich, R.L., Jernigan, M.C., Satterfield, L.C. and Bowles, N.D. (1987) Radiation Carcinogenesis: Time-Dose Relationships. Radia. Res. 111: 179-184.

Underbrink, A.G., Kellerer, A.M., Mills, R.E. and Sparrow, A.H. (1976) Comparison of X-Ray and Gamma-Ray Dose-Response Curves for Pink Somatic Mutations in Tradescantia Clone 02. Rad. and Environ. Biophys. 13: 395-303.

UNSCEAR (1977) Sources and Effects of Ionizing Radiation. Report to the General Assembly, with Annexes. United Nations. NY.

UNSCEAR (1986) Genetic and Somatic Effects of Ionizing Radiation. Report to the General Assembly, with Annexes. United Nations. NY.

Upton, A.C. (1985) Biological Basis for Assessing Carcinogenic Risks of Low-Level Radiation. IN Carcinogenesis. Vol. 10: E. Huberman and S.H. Barr, eds. 381-401. Raven Press. NY.

Upton, A.C. (1986) Historical Perspectives on Radiation Carcinogenesis. IN Radiation Carcinogenesis. 1-10. A.C. Upton, R.E. Albert, F.J. Burns and R.E. Shore, eds. Elsevier. NY.

Upton, A.C. (1987) Cancer Induction and Non-Stochastic Effects. Br. J. Radiol. 60: 1-16.

van der Kogel, A.J. (1979) Late Effects of Radiation on the Spinal Cord. Doctoral Thesis. University of Amsterdam. Radiobiological Inst. of the Organ. for Health Research. The Netherlands.

Van Dyk, J., Mah, K. and Keane, T.J. (1989) Radiation-Induced Lung Damage: Dose-Time-Fractionation Considerations. Radiotherapy and Oncol. 14: 55-69.

Velleman, P.F. and Hoaglin, D.C. (1981) Applications, Basics, and Computing of Exploratory Data Analysis. Duxbury Press. Boston, MA.

Virgo, J.A. (1977) A Statistical Procedure for Evaluating the Importance of Scientific Papers. The Library Quarterly. 47(4): 415-430.

von Essen, C.F. (1960) Roentgen Therapy of Skin and Lip Carcinoma: Factors Influencing Success and Failure. Amer. J. Roentgen. 83: 556-570.

von Essen, C.F. (1963) A Spatial Model of Time Dose Area Relationships in Radiation Therapy. Radiol. 81: 881-883.

von Essen, C.F. (1969) A Practical Time Dose Formula for X-Ray Therapy of Skin Cancer. Br. J. Radiol. 42: 474.

von Essen, C.F. (1972) Clinical Radiation Tolerance of the Skin and Upper Aerodigestive Tract. IN: Frontiers of Radiation Therapy and Oncology. 6: 148-159.

Wald, A. (1947) Sequential Analysis. John Wiley & Sons. NY.

Walinder, G. (1978) Radiation Tumorigenesis in Inbred Laboratory Animals and Cancer Risks in Irradiated Human Populations. IN: Late Biological Effects of Ionizing Radiation. Vol. II. 507-515. IAEA. Vienna.

Walker, A.M. and Rothman, K.J. (1982) Models of Varying Parametric Form in Case-Referent Studies. Amer. J. Epidem. 115(1): 129-137.

Walter, S.D. and Holford, T.R. (1978) Additive, Multiplicative, and Other Models for Disease Risks. Amer. J. Epidem. 108(5): 341-346.

Wara, W.M., Phillips, T.L., Margolis, L.W. and Smith, V. (1973) Radiation Pneumonitis: A New Approach to the Derivation of Time-Dose Factors. Cancer. 32: 547-552.

Weber, N.C. and Welsh, A.H. (1983) Jackknifing and the General Linear Model. Austral. J. Statis. 25(3): 425-436.

Webster, M. (1967) Webster's New International Dictionary.

Weisberg, S. (1985) Applied Linear Regression. 2nd ed. John Wiley & Sons. NY.

Welsch, R.E. (1986) An Introduction to Regression Diagnostics. IN Multiple Regression Analysis: Applications in the Health Sciences. 17-33. D.E. Herbert & R.E. Myers, eds. AAPM Monograph No. 13. AIP. NY.

Weyl, H. (1927/1949) Philosophy of Mathematics and Natural Science. Translated by O. Helmer, 1949. Princeton Univ. Press. Princetnn. 191.

Whewell, W. (1860/1971) On the Philosophy of Discovery. Burt Franklin, NY.

Whittemore, A. and Keller, J.B. (1978) Quantitative Theories of Carcinogenesis. SIAM Rev. 20: 1-30.

Wilks, S.S. (1963) Mathematical Statistics. John Wiley. NY.

Williams, D.A. (1976) Improved Likelihood Ratio Tests for Complete Contingency Tables. Biometrika. 53: 33-37.

Williamson, J.W., Goldschmidt, P.G. and Colton, T. (1986) The Quality of Medical Literature: An Analysis of Validation Assessments. IN Medical Uses of Statistics. J. C. Bailar and F. Mosteller, eds. 370-392. NEJM Books. Waltham, MA.

Withers, H.R. (1986) Predicting Late Normal Tissue Responses. Int. J. Radia. Oncol. Biol. Phys. 2 693-698.

Withers, H.R. (1989) Contrarian Concepts in the Progress of Radiotherapy. Radia. Res. 119. 395-412.

Withers, H.R., Taylor, J.M.G. and Maciejewski, B. (1988a) Treatment Volume and Tissue Tolerance. Int. J. Radia. Oncol. Biol. Phys. 14: 751-759.

Withers, H.R., Taylor, J.M.G. and Maciejewski, B. (1988b) The Hazard of Accelerated Tumor Clonogen Repopulation During Radiotherapy. Acta Oncol. 27: Fasc. 2. 131-146.

Wolins, L. (1962) Responsibility for Raw Data. Amer. Psy. 17: 657-658.

Woodcock, A.E.R. (1978) Catastrophe Theory. 1st ed. NY.

Woolf, P.K. (1988) Deception in Scientific Research. IN Project on Scientific Fraud and Misconduct. 37-86. AAAS-ABA Natl. Conf. of Lawyers and Scientists. Report on Workshop Number One. Sept 18-20, 1987. Hedgesville, WV. AAAS. Washington, DC. 1988.

Wu, C.F.J. (1986) Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. The Annals of Statistics. 14(4): 1261-1295

Yaes, R.J. (1987) Some Implications of the Linear Quadratic Model for Tumor Control Probability. Int. J. Radia. Oncol. Biol. Phys. 14: 147-157.

Yaes, R.J. and Kalend, A. (1988) Local Stem Cell Depletion Model for Radiation Myelitis. Intl. J. Radia. Oncol. Biol. Phys. 14: 1247-1259.

Yankauer, A. (1990) Who are the Peer Reviewers and how much do they Review? JAMA. 263: 1338-1340.

Younger, M.S. (1979) Handbook for Linear Regression. Duxbury Press. North Scituate, MA.

Zelen, M. (1982) Strategy and Alternate Randomized Designs in Cancer Clinical Trials. Cancer Treatment Reports. 66(5): 1095-1100.

Zellner, A. (1971) An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, Inc. NY.

Zellner, A. (1984) Estimation of Functions of Population Means and Regression Coefficients Including Structural Coefficients: A Minimum Expected Loss (MELO) Approach. IN Basic Issues in Econometrics. 238-269. University of Chicago Press. Chicago, IL.

Zellner, A. and Park, S. (1979) Minimum Expected Loss (MELO) Estimators for Functions of Parameters and Structural Coefficients of Econometric Models. J. Amer. Statis. Assn. 74: 185-193.

Ziman, J.M. (1968) Public Knowledge. Cambridge Univ. Press. Cambridge, MA.

Zipf, G.K. (1965) Human Behaviour and The Principle of Least Effort. Hafner Pub. Co. NY.