AbstractID: 3495 Title: Multi-variable modeling of radiotherapy outcomes: determining optimal model size

Purpose: The probability of a specified radiotherapy outcome (e.g., a normal tissue complication or tumor eradication) is typically a complex, non-linear, unknown function of dose distribution characteristics and clinical factors (such as chemotherapy, age, gender, diabetes, etc.). However, current outcome models are usually over-simplified, and standard model fitting methods give little guidance as to how to best add choose from many complicated, alternative models. We discuss methods for building multivariable response models within the framework of logistic regression. We study in detail the issue of how to select model complexity to reach the goal of maximizing predictive power.

Methods: Analyses of esophagitis and xerostomia datasets are used as examples. We describe techniques for approximating the unknown dose-volume-response function as a linear combination of multiple candidate dosimetric variables and clinical factors. In order to guard against under- and over-fitting, we compare several methods for selecting optimal model size, including: fitting against bootstrap training and testing datasets, Akaike information criteria, and leave-one-out cross validation.

Results: Leave-one-out cross validation produced the most unambiguous guidance for optimal model size. Optimal esophagitis model size was five variables (concurrent chemotherapy, A55, A30, A45, A85). Although the xerostomia model could be improved using clinical factors, the improvement over using the single dose-volume model term was small, and therefore judged not worth the added complexity.

Conclusions. Treatment response models, including dose-volume effects, can be made more predictive by mixing clinical and multiple dose-volume factors into a single model. Over-simplified treatment response models are only justified in those cases where more complicated models cannot be supported by the data. Leave-one-out cross correlation model testing combined with Spearman's correlation coefficient often provided the least ambiguous method to study the tradeoff between prediction improvements and model size and to choose optimal model size.