Practical Aspects of CAD Research

Assessment Methodologies for CAD

Charles E. Metz
Professor of Radiology
The University of Chicago

The six levels of diagnostic efficacy:
(Fryback & Thornbury, Med Decis Making, 1991)

1) Technical quality: MTF, NPS, H&D curve, etc.
2) Diagnostic accuracy: Agreement between diagnoses and "truth"
3) Diagnostic-thinking efficacy: Impact of Dx test on physician's thinking about each patient
4) Therapeutic efficacy: Impact of Dx test on patient management
5) Patient-outcome efficacy: Impact of Dx test on patients' health
6) Societal efficacy: Impact of Dx test on society as a whole

Why is receiver operating characteristic (ROC) analysis necessary?
... because of the limitations of other available methods for evaluating diagnostic accuracy

A pair of indices:
“Sensitivity” and “Specificity”

- Sensitivity: Probability of calling an actually-positive case “Positive”
- Specificity: Probability of calling an actually-negative case “Negative”

“Sensitivity” and “Specificity”:

- independent of disease prevalence (if Dx test is used in a constant way)
- implicitly reveal relative frequencies of FP and FN errors

Problems in comparing Dx tests in terms of Sensitivity and Specificity:

- Sensitivity and Specificity of each test depend on the particular "threshold of abnormality" adopted for that test
- Often, one test is found to have higher Sensitivity but lower Specificity than the other
Dependence of Sensitivity and Specificity on “threshold of abnormality”:

\[
\begin{array}{c|c|c}
\text{Specificity} & \text{Sensitivity} \\
\hline
0.0 & 1.0 \\
0.0 & 1.0 \\
\end{array}
\]

A curve is swept out as the “threshold of abnormality” \((t)\) is varied continuously:

An “ROC” curve

The ROC “Area Index” \((A_z)\):

\[
A_z = \begin{cases} 
0.5 & \text{guessing} \\
1.0 & \text{perfect} 
\end{cases}
\]

Interpretations of ROC area \((A_z)\):

- Sensitivity (TPF) averaged over all Specificities (or FPFs) — i.e., average ROC curve height
- Specificity averaged over all Sensitivities
- Probability of distinguishing correctly between a randomly selected actually-positive case and a randomly selected actually-negative case

However ...

- this global index can be misleading when curves cross and/or there is only one region of interest

Other ROC-based indices of performance

- Partial area below, or to the right of, a segment of the ROC curve (regional)
- TPF at fixed FPF or vice-versa (local)
- Expected utility at optimal operating point (local) — most meaningful but least practical

Generalized ROC analysis:

- Localization ROC (LROC) analysis
- Free-response ROC (FROC) analysis
- Alternative FROC (AFROC) analysis
Conventional ROC curves:

- Fraction of actually-positive images falsely called positive (FPF)
- Fraction of actually-negative images falsely called positive (TPF)

LROC (Localization ROC) curves:

- Images must have 0 or 1 lesion
- Curve depends on location-error tolerance

FROC (Free-response ROC) curves:

- Images can have any number of lesions
- Curve depends on location-error tolerance

AFROC (Alternative FROC) curves

- Images can have any number of lesions
- Curve depends on location-error tolerance

The “technology” of ROC analysis:

- Sampling images and readers
- Designing the experiment and collecting observer-response data
- Fitting ROC curves to the data
- Testing the statistical significance of apparent differences between ROC curve estimates

Selecting meaningful samples of cases and readers

- “Absolute measurement” vs. “Ranking” study
  - Absolute measurement: Samples must represent defined clinical populations
  - Ranking: Cases and/or readers can be selected to represent “stressful” subpopulations (e.g., subtle cases and/or expert readers)
    — Generalization of conclusions requires assumptions
- Criteria for inclusion must be explicit
  - Absolute measurement: Define populations sampled
  - Ranking: Report characteristics of cases and readers employed
Designing a study to avoid bias ...

- ... due to absence of subtle disease:
  - Before study is begun, decide criteria for “actually-positive” cases to be included

- ... due to absence of confounding cases:
  - Include clinically-encountered “actually-negative” cases with features that may degrade classifier performance (e.g., cysts in detection of breast cancer)

- ... due to absence of “truth” (verification bias):
  - Establish “truth” for — and include — difficult cases

Avoiding bias in assessments of automated classifiers ...

- ... due to training and testing on same cases:
  - Train and test classifier on different cases subsampled independently from same sample (e.g., “leave-one-out” method)
  - —> Difficult or impossible with rule-based classifiers

- ... due to misinterpretation of meaning and precision of evaluation study’s result:
  - Changing number of training cases changes both true classification accuracy and precision with which true classification accuracy (for a given number of training cases employed) can be estimated
  - Changing number of test cases changes only precision

Avoiding bias in CAD studies...

- ... from failure to consider how CAD will be used:
  - If CAD is to aid human observer, then performance of aided observer must be measured
  - —> Better computer detection scheme may not complement human observer best
  - —> Computer-human interface is crucial

- ... from failure to consider higher-level efficacy:
  - Does/will CAD change patient outcomes?
  - Is/will CAD be cost effective?
  - —> Data are needed — faith is not enough!

Practical issues in designing human observer studies

- Use a continuous or nominally continuous (“100-point”) rating scale
- Use a block design to avoid “reading-order” effects
- In clinical studies, don’t underestimate the difficulty of establishing “truth” without introducing bias

Current controversies:

- Advantages/disadvantages of discrete vs. continuous or nominally continuous (“100-point”) confidence-rating scales?

- Advantages/disadvantages of conventional ROC vs. FROC/AFROC methodology?
  - realism
  - adequacy of information obtained
  - availability of robust curve-fitting and statistical techniques
  - statistical power

The “technology” of ROC analysis:

- Sampling images and readers
- Designing the experiment and collecting observer-response data
  - Fitting ROC curves to the data
  - Testing the statistical significance of apparent differences between ROC curve estimates
**ROC curve fitting**

- Some functional form with adjustable parameters must be assumed for the ROC curve — usually the “binormal” model.
- The assumptions of conventional least-squares curve fitting aren’t valid here, so maximum-likelihood (ML) estimation should be used instead.
- Free software is available (listed later).

**ROC curve fitting (continued)**

The conventional “binormal” curve-fitting model ...

- Assumes that all ROC curves plot as straight lines on “normal deviate” axes ($z_{TPF}$ vs. $z_{FPF}$).
- Equivalently, assumes that the two underlying distributions can be transformed to normal by a generally unknown transformation (“semi-parametric”).
- Has been shown valid in a broad variety of situations but ...
- Can yield inappropriate shapes when cases are few and/or when data scale is discrete and operating points are poorly-distributed (→ “proper” models).

**Statistical significance tests for differences between ROC curves**

Ways that “difference” can be quantified:

- Area index $A_z$ (global)
- TPF at a given FPF (local)
- FPF at a given TPF (local)
- Partial area index (“regional”)
- Both parameters of binormal model (“bivariate”)
- Cost/Benefit (at optimal operating points)

**Statistical significance tests (cont’)**

Different statistical tests take different kinds of variation taken into account (and, thus, allow different generalizations):

- **Reader variation only** (a “significant” result applies to readers in general … but only to the particular cases used in the experiment)
- **Case-sample variation only** (result applies to cases in general … but only to the particular reader[s] used)
- Both (result applies to readers and cases in general)

⇒ Note: Conventional statistical tests cannot be applied directly in most situations.

**Current statistical tests …**

… that take only reader variation into account:

- paired or unpaired Student’s $t$ test of differences in any index … at least in principle

**Current statistical tests…**

… that take only case-sample variation into account:

- non-parametric Wilcoxon/Mann-Whitney tests of differences in total ROC area [only] (Hanley & McNeil, DeLong et al.)
- non-parametric tests of differences in any index … at least in principle (Wieand et al.)
- semi-parametric tests of differences in any index … at least in principle (Metz et al.)
Current statistical tests...

... that take both sources of variation into account (and are applicable to differences in any index, at least in principle):

- semi-parametric tests (Swets & Pickett; Dorfman, Berbaum & Metz; Tolédano & Gatsonis; Obuchowski)
- bootstrapping approach (Beiden, Wagner & Campbell)

Free software for ROC analysis:

- Metz (University of Chicago; >5000 registered users)
  - ROCFIT and LABROC: fits a single ROC using the binormal model
  - INDRROC: tests difference between independent ROC estimates
  - CORROC2 and CLABROC: test diff. between correlated ROCs
    - difference in $A_z$
    - difference in TPF at given FPF
    - diff. in both binormal ROC curve parameters ("bivariate" test)
  - ROCKIT: integrates and extends the five programs above
- Dorfman and Berbaum (University of Iowa)
  - RSCORE2 and RSCORE4: fit a single ROC using binormal model
  - MRMC: Jackknife-ANOVA test for difference in $A_z$ (data collected on continuous and discrete scale)

All University of Chicago software for ROC curve fitting and statistical testing can be downloaded from the World Wide Web without charge from:

http://xray.bsd.uchicago.edu/krl/roc_soft.htm

>>> Please note new URL

Current controversies:

- Best way to fit ROC curves to “degenerate” data?
  - RSCORE4 (ad hoc)
  - bigamma model (restricts curve shape too much?)
  - “proper” binormal model (computationally intensive, no statistical tests for differences so far)
  - “contaminated” binormal model (restricts curve shape too little?)
- Validity/robustness of current techniques for fitting FROC/AFROC curves and testing the statistical significance of differences thereof?
- Most appropriate index/indices for comparisons?

Relationship between ROC analysis and Cost/Benefit analysis:

- Different “operating points” on an ROC curve provide different frequencies of TP, FP, TN, and FN decisions (which depend on disease prevalence).
- If utilities can be assigned to the various kinds of correct and incorrect decisions and if prevalence is known, then the optimal operating point can be found on any ROC curve.
- The maximized utility found in this way quantifies the “value” of a diagnostic test in terms of its ROC.
- See reading list for details.

Needs for the future:

- Develop stratified-sampling methodology
- Establish validity/robustness of data-analysis techniques for free-response paradigms
  - curve fitting
  - statistical testing of differences
- Develop “MRMC” methods for statistical analysis of data from incompletely-balanced experimental designs, particularly...
  - when observers don’t read the same cases
  - when data are correlated within cases
Needs for the future (continued):

- Develop highly efficient approaches well-suited to exploratory analyses
  - Key need is to control for decision-threshold effects
  - Other biases may be acceptable if sufficiently small
- Generalize ROC analysis to handle >2 decision alternatives
  - Must provide an appropriate compromise between complexity and practicality
- Approaches proposed to date are not adequate

An incomplete list of recommended literature on ROC methodology

- **BACKGROUND:**

- **CURVE FITTING:**

- **BIAS:**
  - Swets JA. Indices of discrimination or diagnostic accuracy: their ROCs and implicit models. Psychol Bull 1986; 99: 149.

- **STATISTICS:**
  - Begg CB, Greensa RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 151.
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1982; 143: 266.