DOE/ER/61816-1

US Department of Energy Grant DE-FG02-94ER61816

"Methodology for Evaluation of Diagnostic Performance"

Funding period: 4/1/97 - 3/31/99

Charles E. Metz, Ph.D., Principal Investigator
Department of Radiology
The University of Chicago

Final Technical Report

A. Overview

The proliferation of expensive technology in diagnostic medicine demands objective, meaningful assessments of diagnostic performance. Receiver Operating Characteristic (ROC) analysis is now recognized widely as the best approach to the task of measuring and specifying diagnostic accuracy (Metz, 1978; Swets and Pickett, 1982; Beck and Schultz, 1986; Metz, 1986; Hanley, 1989; Zweig and Campbell, 1993), which is defined as the extent to which diagnoses agree with actual states of health or disease (Fryback and Thornbury, 1991; National Council on Radiation Protection and Measurements, 1995). The primary advantage of ROC analysis over alternative methodologies is that it separates differences among diagnostic decisions that are due to actual differences in discrimination capacity from those that are due to decision-threshold effects (e.g., "under-reading" or "over-reading"). An ROC curve measures diagnostic accuracy by displaying True Positive Fraction (TPF: the fraction of patients actually having the disease in question that is diagnosed correctly as "positive") as a function of False Positive Fraction (FPF: the fraction of patients actually without the disease that is diagnosed incorrectly as "positive"). Different points on the ROC curve — i.e., different compromises between the specificity and the sensitivity of a diagnostic test, for a given inherent accuracy — can be achieved by adopting different critical values of the diagnostic test's "decision variable " — e.g., the observer's degree of confidence that each case is positive or negative in a diagnostic image-reading task, or the numerical value of the result of a quantitative diagnostic test.

ROC techniques have been used to measure and specify the diagnostic performance of medical imaging systems since the early 1970s, and the needs that arise in this application have spurred a variety of new methodological developments. In particular, substantial progress has been made in ROC curve fitting and in developing statistical tests to evaluate the significance of measured differences between ROC curves. These are especially important tasks in medical applications, because various practical issues usually limit the number of patients with clearly established diagnostic truth that can be included in any study that seeks to measure diagnostic performance objectively. Other progress has been made in relating ROC analysis to cost/benefit analysis, and in

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

generalizing ROC methods to accommodate some diagnostic tasks where more than two decision alternatives are available.

ROC analysis clearly provides the most rigorous and fruitful approach for such assessments but, like many other powerful techniques that provide useful insight concerning complex situations, it currently suffers from limitations, particularly in evaluation studies that involve small case samples. However, the potential of this relatively new analytic approach and the concepts on which it is based have not been fully explored. The research proposed here is designed to refine and supplement existing ROC methodology to increase both the accuracy and the precision of its results.

Dr. Metz has played a key role since the early 1970s in developing ROC methodology for the evaluation of diagnostic medical procedures. He and his colleagues were the first to generalize ROC analysis so that it applies to diagnostic tasks involving more than two decision alternatives (Starr, Metz, Lusted and Goodenough, 1975; Metz, Starr and Lusted, 1976); to propose and validate a formal statistical test for differences between binormal ROC curve estimates (Metz and Kronman, 1980); to propose and validate a parametric approach for testing the significance of differences between ROC curves estimated from correlated data (Metz, Wang and Kronman, 1984); and to propose and validate a model that predicts the gains in accuracy which are available from replicated readings of diagnostic images (Metz and Shen, 1992). Dr. Metz's tutorial publications (especially Metz, 1978; Metz, 1986; and Metz, 1989) also have fostered the now-widespread use of ROC analysis in medical imaging and its growing acceptance in other medical disciplines.

B. Research accomplishments

1. Distribution of ROC software

We have been developing computer software for maximum-likelihood estimation of ROC curves and for testing the statistical significance of differences between ROC curves and indices since the late 1970s (Metz and Kronman, 1980; Metz, Wang and Kronman, 1984; Metz, 1989). Currently, eight programs are available in versions for computers that employ the Microsoft Windows and UNIX operating systems. The six most frequently requested programs are available also in versions for the Apple Macintosh.

Our software, which we provide without charge to all investigators who request it, is now generally accepted as the standard for ROC data analysis in medical imaging applications. Most of our software's users obtain the programs by downloading them from a World Wide Web page which we created for that purpose at <http://www-radiology.uchicago.edu/cgi-bin/software.cgi>. Beginning with 667 registered users on April 1, 1997, an additional 836 registered users obtained copies during the subsequent two-year funding period, thereby achieving a total of 1503 registered users on March 31,

1999. On January 7, 2003, as this report is written, the total number of registered users has reached 5667.

2. Maximum-likelihood estimation of ROC curves from continuously-distributed data

Until recently, rigorous ROC curve-fitting techniques were available only for data collected on a discrete categorical scale. However, data from clinical laboratory tests usually are collected on a continuous scale, and continuous scales are used increasingly to collect data in diagnostic image-reading studies (Rockette, Gur and Metz, 1992). We were able to prove theoretically that "truth-state runs" in rank-ordered outcomes of continuously-distributed data constitute a natural "categorization" of such data for maximum-likelihood (ML) estimation of ROC curves. On the basis of this insight, we developed two new algorithms for fitting binormal ROC curves to continuously-distributed data: a true ML algorithm (LABROC4) and a quasi-ML algorithm (LABROC5) that requires substantially less computation with large datasets. Extensive simulation studies demonstrate that both algorithms produce reliable estimates of the binormal ROC parameters *a* and *b*, the ROC-area index A_Z, and the standard errors of those estimates. A paper describing our approach and the LABROC algorithms (Metz, Herman and Shen, 1998) provided the first firm theoretical basis for fitting ROC curves to continuously-distributed data by maximum likelihood estimation, though an alternative, computationally more demanding approach based solely on rank-order statistics was proposed subsequently by Zou and Hall (1997).

3. Maximum-likelihood estimation of "proper" binormal ROC curves

The conventional binormal model of ROC analysis — which assumes that a latent (i.e., effective) pair of normal decision-variable distributions can be used to represent experimental data but does *not* assume that the data themselves are normally distributed (Metz, 1986; Hanley, 1988) — has been used for many years to fit smooth ROC curves to data (Dorfman and Alf, 1969; Swets, 1986; Hanley, 1988; Hanley, 1996; Hajian-Tilaki *et al.*, 1997), and algorithms such as RSCORE (Donald D. Dorfman, Ph.D., in Swets and Pickett, 1982) and ROCFIT (Metz, 1989) are readily available for maximum-likelihood estimation (MLE) of such curves. When ROC curves with this form are fit to sufficiently large datasets, they rise rapidly from the lower-left corner of the unit square and then bend smoothly and steadily into the upper-right corner. However, if the conventional binormal model is used for small datasets or datasets with poorly allocated category boundaries, a "hook" in the fitted ROC curve may be evident near the upper-right or lower-left corner of the unit square, causing the neighboring part of the ROC to drop below the 45° "guessing line." Such ROC curves are said to be "improper," because their non-monotonic slope indicates that they could not have been produced by an optimal decision rule. In extreme situations of this kind, the data are fit exactly by a "degenerate" limiting form of the conventional binormal ROC that consists of vertical and horizontal line segments (Metz, 1989). To overcome these curve-fitting artifacts, we developed a "proper" binormal model and a new algorithm for MLE of the corresponding ROC curves. Like the conventional binormal model, our new model is based upon an implicit assumption that an effective pair of normal distributions underlies the data. However, the "proper" binormal model assumes that the ROC data were produced by a decision variable that corresponds to the likelihood ratio associated with the pair of normal distributions, rather than to the normally-distributed quantity itself.

MLE of the parameters of the proper binormal model can be difficult, because some substantially different combinations of parameter values produce very similar ROCs, causing the likelihood functions of some datasets to have long, narrow "ridges" in parameter space. We were able to overcome this problem by appropriate re-parameterization of the model and use of a novel iterative scheme. Extensive simulation studies have shown the resulting curve-fitting algorithm, entitled PROPROC, to be highly robust. Maximum-likelihood ROC curve estimates obtained from the proper and conventional binormal models were virtually identical when the conventional binormal ROC showed no "hook," but the proper binormal curves had monotonic slope for all datasets, including those for which the conventional model produced degenerate (Metz, 1989) fits. Our simulation studies showed also that PROPROC estimated the population ROC curves and the total-area (Metz, 1986) and partial-area accuracy indices (McClish, 1989; Jiang, Metz and Nishikawa, 1996) of the population ROC curves with little bias. We published this work in two papers, one that focuses primarily on the ability of the proper-binormal approach to deal with "degenerate" datasets (Pan and Metz, 1997) and another that provided a detailed description of the theory and a computational algorithm that implements it (Metz and Pan, 1999). We also collaborated in developing and evaluating an alternative "proper" ROC approach that is based upon an underlying pair of gamma distributions (Dorfman, *et al.*, 1997).

4. Statistical comparison of two ROC curve estimates obtained from partially-paired datasets

All previously-available techniques for testing the statistical significance of differences between ROC curve estimates apply only to <u>either</u>: (i) "fully-paired" datasets, in which both diagnostic modalities in a comparison are applied to all of the patients in a single case sample; <u>or</u> (ii) "unpaired" datasets, in which wholly independent case samples are obtained for the two modalities. Many evaluation studies are designed to obtain fully-paired datasets, due to the greater statistical power that pairing endows, but in practice data from only a single modality often are obtained for some patients. Stimulated by this need, we completed development of a rigorous theoretical basis for ROC analysis of such "partially-paired" datasets, and we developed, tested and debugged a new generalization of our CORROC algorithm (Metz, Wang and Kronman, 1984) that applies to such "partially-paired" datasets. We then performed extensive computer-simulation studies that evaluated the validity of the new algorithm's statistical test for differences in the conventional binormal ROC area index, A_z, by investigating the relationship between the algorithm's empirical Type I error rate and the critical p-value (α) used for the test. We described this work in detail in a paper published in *Medical Decision Making* (Metz, Herman and Roe, 1998).

5. ROCKIT: an integrated software package for analysis of ROC data

We expect to release soon a new software package entitled ROCKIT that integrates five of our current programs (ROCFIT, LABROC5, INDROC, CORROC2, and CLABROC) into a single shell that will: (i) fit a single conventional binormal ROC curve to a single set of data collected on either a discrete or a continuous scale; (ii) test the statistical significance of a difference between conventional binormal ROC curves estimated from either paired or unpaired data that were collected on discrete and/or continuous scales; and/or (iii) estimate the effective correlation of paired data that are collected on discrete and/or continuous scales. An important innovation in ROCKIT is its ability to test the statistical significance of differences between ROC curves that are estimated from *partially*-paired case samples, as described immediately above (Section 4). ROCKIT also is able to read input files created by any of the five existing programs that it integrates, and it outputs 95% confidence intervals (Metz, 1993) for all of the estimates that it provides.

6. Interpretation of variance components in analyses of ROC data

Multivariate linear ("ANOVA") models are being used increasingly in ROC analysis to assess statistical variation in ROC estimates (Dorfman, Berbaum and Metz, 1992; Metz and Shen, 1992; Beam, 1995; Dorfman and Metz, 1995; Gatsonis, 1995; Obuchowski, 1995). Most such models proposed for ROC applications represent the overall statistical variation in latent decision variables and/or in ROC index estimates by a sum of uncorrelated components that can be ascribed to differences between the diagnostic modalities under investigation, case-sample variation, reader variation, and interactions among these factors. However, the abstract nature of this practically-important statistical approach seems to have intimidated many investigators, thereby slowing its widespread adoption. In an attempt to overcome this problem, we developed notation that readily distinguishes variances and correlations that are associated with each method of replication and, for estimate differences, each estimate-pairing scheme that might be used in an ROC experiment. We then considered a general variance-component model for ROC index estimates and differences thereof, and we used that model to systematize the many variances and correlations that are observable in ROC experiments. In this way we were able to demonstrate hopefully intuitive relationships among notation, different methods of experimental replication, and particular components of the multivariate linear model.

Specifically, we delineated four methods of replication and eight pairing schemes for generating ROC index-estimate differences by using a mixed linear model with one fixed factor (modality) and two random factors (reader and case sample). For each of the resulting 32 replication-pairing combinations, we then systematically expressed the variance of the difference and the correlation between the two ROC estimates in terms of the variance components of our model (Roe and Metz, 1997). After exploring the relationship between expressions derived from our general multivariate linear model and expressions given by Swets and Pickett (1982), we concluded that the Swets/Pickett approach overestimates variance in some situations (Roe and Metz, 1997).

C. Literature Cited

Beam CA (1995). Random-effects models and the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches and issues. *Academic Radiol.* **2** (Supplement 1): S4-S13.

Beck JR and Shultz EK (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch. Pathol. Lab. Med.* **110**: 13-20.

Dorfman DD, Berbaum KS and Metz CE (1992). ROC rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest. Radiol.* **27**: 723-731.

Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA and Dagga HA (1997). Proper ROC analysis: the bigamma model. *Academic Radiol.* **4**: 138-149.

Dorfman DD and Metz CE (1995). Multi-reader multi-case ROC analysis: comments on Begg's commentary. *Academic Radiol.* **2** (Supplement 1): S76-S78.

Fryback DG and Thornbury JR (1991). The efficacy of diagnostic imaging. *Med. Decis. Making* **11**: 88-94.

Gatsonis CA (1995). Random-effects models for diagnostic accuracy data. *Academic Radiol.* **2** (Supplement 1): S14-S21.

Hajian-Tilaki KO, Hanley JA, Joseph L and Collet J -P (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Med. Decis. Making* **17**: 94-102.

Hanley JA (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Med. Decis. Making* **8**: 197-203.

Hanley JA (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging* **29**: 307-335.

Hanley JA (1996). The use of the "binormal" model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* **15**: 1575-1585.

Jiang Y, Metz CE and Nishikawa RM (1996). An ROC partial area index for highly sensitive diagnostic tests. *Radiology* **201**: 745-750.

McClish DK (1989). Analyzing a portion of the ROC curve. *Med. Decis. Making* **9**: 190-195.

Metz CE (1978). Basic principles of ROC analysis. *Seminars Nucl. Med.* **8**: 283-298.

Metz CE (1986). ROC methodology in radiologic imaging. *Invest. Radiol.* **21**: 720-733.

Metz CE (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol.* **24**: 234-245.

Metz CE (1993). Quantification of failure to demonstrate statistical significance: the usefulness of confidence intervals. *Invest. Radiol.* **28**: 59-63.

Metz CE, Herman BA, Roe CA (1998). Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Med. Decis. Making* 18: 110-121, 1998.

Metz CE, Herman BA, Shen J-H (1998). Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Statist. Med.* 17: 1033-1053.

Metz CE and Kronman HB (1980). Statistical significance tests for binormal ROC curves.
J. Math. Psychol. **22**: 218-243.

Metz CE and Pan X (1999). "Proper" binormal ROC curves: theory and maximum-likelihood estimation. J. Math. Psychol. 43: 1-33.

Metz CE and Shen J-H (1992). Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. *Med. Decis. Making* **12**: 60-75.

Metz CE, Starr SJ and Lusted LB (1976). Observer performance in detecting multiple radiographic signals: prediction and analysis using a generalized ROC approach. *Radiology* **121**: 337-347.

Metz CE, Wang P-L and Kronman HB (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In: *Information Processing in Medical Imaging* (F Deconinck, ed.). The Hague: Nijhoff, pp. 432-445.

Obuchowski NA (1995). Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Academic Radiol.* **2** (Supplement 1): S22-S29.

National Council on Radiation Protection and Measurements (1995). *An Introduction to Efficacy in Diagnostic Radiology and Nuclear Medicine* (NCRP Commentary No. 13). Bethesda, Maryland: NCRP.

Pan X and Metz CE (1997). The "Proper" binormal ROC model: parametric estimation of degenerate ROC data. *Academic Radiol.* **4**: 380-389.

Rockette HE, Gur D and Metz CE (1992). The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Invest. Radiol.* **27**: 169-172.

Roe CA and Metz CE (1997). Variance-component modeling in the analysis of ROC index estimates. *Academic Radiol.* (in press).

Starr SJ, Metz CE, Lusted LB and Goodenough DJ (1975). Visual detection and localization of radiographic images. *Radiology* **116**: 533-538.

Swets JA and Pickett RM (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

Zou KH and Hall WJ (1997). Semiparametric and parametric transformation models for estimating a receiver operating characteristic (ROC) curve from continuous diagnostic test data. Technical Report 97/01, Department of Biostatistics, University of Rochester, 1997. Presented at the 1997 ENAR Spring Meeting, Memphis, TN, 1997.

Zweig MH and Campbell G (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin.Chemistry* **39**: 561-577.