Recent Developments and Contemporary Issues for Diagnostic Technology Assessment

Robert F. Wagner, Ph.D.
Center for Devices & Radiological Health/FDA
Rockville MD 20850

ABSTRACT

This paper summarizes recent themes of interest to this conference as well as the conferences of the Medical Image Perception Society (MIPS).

Keywords: ROC analysis, reader variability, MRMC ROC, ACRIN DMIST, ROC software uprgrades, multiple-biomarkers, cross-validation, bootstrapping, JAFROC

1. BACKGROUND

I wish to thank our chairman and the organizers of this conference for the invitation to address this gathering of many of our distinguished colleagues and friends. We represent a community that has been developing a culture whose purpose is the assessment of diagnostic technologies using finite (i.e., modest or "least burdensome") resources. Several of us have recently attempted to summarize the state of consensus development toward this goal. The first major paper in this series reviewed some public experience and case histories of submissions to the FDA in the area of digital mammography and computer-aided diagnosis in breast and lung imaging (1). A follow-up paper presented at one of our SPIE Medical Imaging Conferences provided some practical rules of thumb for investigators involved in designing studies in these areas (2). A PowerPoint presentation at an FDA Radiology Panel Meeting displayed some of the issues involved in extending this work to the area of lung nodule detection on CT scans (3). We are currently in the process of putting together an up-dated version of some of these materials (4). The purpose of this presentation today is to highlight some topics being formulated for the up-date, with emphasis on recent developments. Definitions of the terms used below can be found in the references just provided.

2. THE KEY POINT OF ROC ANALYSIS

A comparison of two modalities in terms of a single sensitivity-specificity (Se/Sp) point for one modality versus that for the competing modality is only unambiguous if one of the modalities surpasses the other in both sensitivity and specificity. (A slight extension of this point for the case of clinical laboratory tests has been given by Biggerstaff [5]). Otherwise, there is a trade-off between the two modalities which requires specifying a utility matrix and carrying out what is called "expected benefit" or "expected utility" analysis at those particular Se/Sp points [6]. On the other hand, if one measures the entire receiver operating characteristic (ROC) curve for both modalities, one can compare the full trajectory of the Se/Sp of one modality versus that of the other and the ambiguity can generally be resolved without struggling with the difficult specification of utilities. The most popular introduction to these issues was given by Metz in 1978 [7].

3. SCALES FOR MEASURING ROC CURVES

ROC analysis involves making measurements on observers of images and then analyzing the resulting data. The first or historic measurement scale for collecting such data was implemented by requesting the image reader to

Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment, edited by Yulei Jiang, Miguel P. Eckstein, Proc. of SPIE Vol. 6146, 614606, (2006) · 1605-7422/06/$15 · doi: 10.1117/12.663460

report on a five- (or six- or seven-) category rating scale, ranging from "almost certainly normal" to "almost certainly abnormal"—through a set of intermediate categories. This represented a breakthrough from the earliest ROC studies which required repeating the experiment multiple times with different utilities (or "payoffs") and costs (or penalties) to the readers to get them to adopt different levels of aggressiveness toward making positive calls, and thus move them along their ROC curves. Once the concept and use of the rating scale was introduced, it became possible to obtain the data required for an ROC study on a single pass. The curve was generated by moving a threshold through the ratings – in the same manner as if the ratings were the output of a laboratory diagnostic test, e.g., the PSA test for suspicion of prostate cancer, and the Se/Sp generated as a function of the threshold just as is done for such laboratory tests.

Maximum-likelihood methods of curve fitting were developed to interpolate and extrapolate such data with smooth parametric fits. To avoid some difficulties associated with convergence of such fits, investigators suggested using a quasi-continuous or 100-point probability-of-disease scale, often implemented by using a track-ball on a computer display of such a scale. (It is curious that this is reminiscent of early psychophysical studies in which a subject would make a mark along a ruler or other visual analog scale.)

Finally, the patient-management—or so-called "action item"—reporting scale appears to be the most familiar scale for clinicians. Such a scale distinguishes only between "no action" versus "follow up" or between "follow up" and "biopsy." The BIRADS categories for mammography [8] represent such a scale. An often missed point is that the BIRADS document actually recommends that the radiologist provide to the referring physician an estimate of the probability of cancer. Thus, the sometimes heard objection that a probability scale is not a practical clinical one does not stand up against the recent historical record.

4. THE HISTORIC NCI CONTRACT FOR AN ROC PROTOCOL

The revolution in imaging technology made possible with the advent of small computers in the 1970s had the dual results of great increases in diagnostic information from new imaging modalities (e.g., CT and MRI) as well as the concomitant increase in the cost of the technologies. This led the National Cancer Institute (NCI) to write a contract for the development of a formal protocol for the collection and analysis of data for the purposes of ROC analysis. The classic book by Swets and Pickett documents the evolution of this work [9]. The Advisory Panel gathered for this project included S. James Adelstein, Harold L. Kundel, Lee B. Lusted Barbara J. McNeil, Charles E. Metz, and J.E. Keith Smith. The emergence of the contemporary state of our field cannot be appreciated without referring to this work. We have the opportunity to receive some of the history of this period first-hand from Harold L. Kundel, M.D., a member of that Advisory Panel who is a frequent contributor to these SPIE Conferences.

5. THE COMPLICATION OF READER VARIABILITY

In the last decade, one study after another has documented the existence of a wide range of level of performance—i.e., reader skill—in studies of human observers using medical imaging systems. This additional variable complicates the problem of ROC analysis. The first formal and practical method for incorporating the uncertainty in performance assessment that arises from the range of reader level of performance was provided by Dorfman, Berbaum, Metz [10], the so-called DBM method. This more general ROC paradigm is often referred to as the multiple-reader, multiple-case (MRMC) ROC paradigm. Many of the early key references in this field are given in the first few references of the present overview [1-3].

6. CONTEMPORARY DEVELOPMENTS IN MRMC ROC ANALYSIS

The DBM analysis of MRMC ROC data uses the jackknife resampling strategy in combination with a linear components-of-variance model and methods of inference based on multivariate-normal-based ANOVA to produce confidence intervals on the difference of mean performance across competing diagnostic modalities. A somewhat different formulation and solution was soon presented by Obuchowski and Rockette [OR: 11, 12]. Hillis and colleagues

[13] made a discovery—somewhat surprising to devotees of one or the other approach—that the DBM and OR procedures yield identical test statistics under certain common conditions, but can lead to considerably different inferences due to their different methods of estimating the degrees of freedom of the denominator of the T^2 (or F) test statistic used in these approaches. Hillis et al. characterized the DBM model as a "working model" of the more fundamental OR conceptualization.

Hillis and Berbaum [14] and Hillis [15] made further developments that lead to what they call "new DBM"—presented at the Medical Image Perception Society (MIPS) Conference in Windermere UK in 2005 [16]. These include a renormalization of the jackknife procedure, less data-driven testing of interaction terms in the denominator of the test statistic, and a fresh approach to estimating the degrees of freedom of the test statistic. These modifications appear to almost eliminate the sometimes conservative inferences of the original DBM approach and offer the same or more statistical power for the same expenditure of resources. The latter follows from the equal or greater number of degrees of freedom calculated in the new DBM method compared to that from the original DBM or OR methods together with the fact that the method has been validated over many variance-covariance structures.

Barrett, Kupinski, and Clarkson [BKC: 17, 18] have recently examined the MRMC ROC problem from the point of view of the fundamentals of probability theory, without the need for invoking any particular components-of-variance model, linear or otherwise. They also refer to their approach as a "mechanistic" one for the following reason. If one can define an imaging system and a method for training a sample of readers of the images (including, e.g., drawing a sample of image training cases from a specified ensemble or population of cases) and drawing a sample of test cases (e.g., from a specified ensemble, presumably the same as that from which the trainers are drawn), then one has completely specified the experiment and can study its statistical properties over Monte Carlo trials—without the need to invoke an abstract variance-covariance structure as used in the validations of the DBM approach. At the present moment, their approach depends on using resampling methods to obtain practical results. Also, they have so far only estimated variances; no inferential machinery, i.e., method for calculating C.I.s, has been proposed. Nevertheless, their approach appears to be a fundamental reformulation of this central problem of our time and is open-ended with respect to future developments.

Gallas has presented a variation on the BCK approach, which he refers to as a "one-shot" solution [19, 20], for the case where the non-parametric Mann-Whitney-Wilcoxon statistic is used as the summary accuracy measure. He uses the expression "one-shot" to indicate that no resampling is required to implement his method. It is thus a multiple-reader generalization of the approach presented by Campbell, Douglas, Bailey [21] for the single-reader problem. Also, in the spirit of U-statistics, his solution is unbiased by construction. Gallas applied his approach to a family of twenty-seven experiments that involved finite training of model observers; the model observers were both noisy (from an internal noise source) and variously handicapped (by denying them access to a fraction of the image pixels). His results are unbiased, as expected; the DBM analysis displayed a small bias for two or three of the experiments; however, the two-way bootstrapping (cases and readers) used by Beiden, Wagner, Campbell [22] displayed very large biases in about half of his experiments, but almost no bias in the other half. At last year's SPIE Medical Imaging Conference, Samuelson and Wagner [23] showed that the bias observed in straightforward two-way bootstrapping could be remedied by a second-order approach to obtaining bootstrap confidence intervals, namely, bootstrap-t confidence intervals.

Bandos, Rockette, Gur [24] have recently demonstrated for the MRMC ROC problem a formulation of the *ideal bootstrap*—a method of obtaining statistical moments of an accuracy measure directly from the empirical distribution, without the need to do statistical resampling. Their formulation allows them to identify the nature of the bias in the two-way bootstrap in the context of the linear model. Their surprising result shows that the bias in the simple two-way bootstrap persists asymptotically as the number of readers grows without bound. Gallas has found an analogous result in his one-shot treatment that does not require the linear model.

A helpful by-product of the analysis by Gallas was a demonstration that the lowest-order terms of the BKC approach line up with the results that are obtained when using the linear components-of-variance model. However, additional terms appear in the former treatment from differences across the normal and abnormal classes of patients, as well as several higher-order terms in the numbers of reader and case samples.

7. THE ACRIN DMIST STUDY

Results from the ACRIN Digital Mammographic Imaging Screening Trial (DMIST) are now starting to be published [25]. This study shows the similarity of digital and conventional mammography when the entire cohort of patients is included, but also the great superiority of digital over conventional mammography for particular subsets of subjects (e.g., those with dense or heterogeneous breasts). Several features of their displayed results are of great interest to our community.

The most obvious feature of their figures is the presence of a hook near the upper-right-hand corner of ROC space for all of the fitted ROC curves for conventional mammography. Although it is physically possible that such hooks can arise from human or animal ROC studies, this phenomenon is more typically a result of the traditional binormal maximum-likelihood ROC curve-fitting algorithm; this algorithm allows such "Improper ROC" curves, i.e., where the slope of the curve does not change monotonically. It is possible that the existence of hooks in the present analysis may slightly enhance the significance of the differences inferred across modalities. We return to the issue of proper and improper ROC curves presently.

Another feature of the results is that the "empirical" or observed points in ROC space were obtained using a seven-category rating scale, and the non-trivial "interior points" fell between false-positive rates of zero and about twenty-five percent. Moreover, the use of BIRADS categories to set the threshold for action in this study led to a false-positive fraction in the neighborhood of 10%. Such a result is consistent with the public-health philosophy that a screening modality should operate with a small false-positive fraction so as not to deter compliance with a national screening program. The other low false-positive results from categorical ratings are also something that might be expected in a screening trial. Such results—limited to only a fraction of ROC space—call out to be analyzed using a measure of the partial-area under the ROC curve as the summary index, as suggested by Baker and Pinsky [26] for the case of screening studies, rather than the total area under the ROC curve.

A curiosity in these results that many (if not most) investigators have not observed before is the way in which the false-positive fractions corresponding to the categorical ratings line up almost exactly for the competing modalities—which then differ only in true-positive fraction—allowing for great simplification in the analysis of the latter. While it is expected that the great number of normal cases will result in very high precision for the estimated false-positive fractions, there is no prior reason to expect the results across modalities to line up vertically as in this study. A large number of readers were used in the study. A given reader only read a given case once; and the digital and the conventional mammograms were read by different readers. In this context, it seems remarkable that the average use of the category ratings turned out to be so consistent across modalities. This is another issue that calls out to be understood at a more detailed level.

A satisfying result of this study is the fact that the BIRADS-based patient-management threshold always yielded Se/Sp points that fell almost on top of the ROC curves that were fitted to the seven-category data. We return to this point below.

Finally, note that the present context is referred to as that of the "pooled reader" or the "generic reader." That is, the rating data is pooled and used as if it were due to a single reader and analyzed using the University of Chicago software CORROC. There is no inferential algorithm or corresponding software to date that includes the variability of the readers over and above that of the cases in the analysis of such results.

8. AVAILABILITY OF RECENT UPGRADES IN ROC SOFTWARE

The following information comes from a recent personal communication [27]. Software has been developed and validated for maximum-likelihood estimation of "proper" binormal ROC curves [28], which do not allow hooks but otherwise provide the same fits as the conventional binormal model. This software, referred to as PROPROC:

- has been tested extensively with simulated and real data;

- allows the use of total-area *or partial*-area accuracy indices;

- now exists in a new version of the University of Iowa/University of Chicago MRMC/LABMRMC software for DBM MRMC analysis that incorporates PROPROC, and should be available now from the University of Iowa website [29] (and soon from the University of Chicago website [30]);

- and a standalone version of PROPROC (analogous to the traditional ROCFIT) will be available soon from the University of Chicago website.

9. A RECENT EDITORIAL

The consistency observed above between the BIRADS-based Se/Sp points and the ROC curves fitted to the seven-category data in the ACRIN DMIST study may require special training of readers, whether in that particular application or further applications where different versions of a categorical scale or a quasi-continuous probability scale or action-item scale are used. (Note that the effectively seven-category new BIRADS rating scale—which expands category 4 to 4a, 4b, 4c—had not been published prior to the DMIST study. Thus, there was no explicit attempt to line up their seven-category rating scale with a BIRADS categorization.) A recent editorial [31] calls on professional and academic societies as well as interested industry and government bodies to work toward consensus and training toward such consistency to strengthen both the science and the statistical power of such studies. This becomes even more important as we move toward an era in which multiple biomarkers will be fused by computer and combined quantitatively with human observer rating data both for diagnostic and prognostic classification of patients.

10. COMPUTER-AIDED DETECTION

Current studies of the use of computers to cue suspicious regions on images—the so-called "spell checker" mode or computer-aided detection (CADe)—often are complicated by the fact that the computer-aided human reader gains in sensitivity but possibly at a small increase in false-positive fraction. An in-depth "effective benefit" analysis of this trade-off requires very large studies as well as historic databases of outcomes to provide the input to the utility matrix. The contemporary holding pattern has been to carry out a modest "stress test"—using a set of images enriched with the target of interest (lung nodules, breast lesions, colon polyps, etc.) as a surrogate for a large realistic clinical "field study." In such stress studies, one collects probability-rating data to generate the full ROC curves of readers unaided versus aided by the computer. In these studies it may be necessary to concentrate on training readers to use the scales of particular interest for the study in a self-consistent way, as emphasized in earlier paragraphs here. It may also be important to elaborate methods for encouraging readers to provide readings that will fill up more of ROC space than was exhibited in the ACRIN DMIST study referred to above.

11. COMPUTER DIAGNOSIS

We are crossing the threshold of an era in which computers will be used not only as aids to readers but as primary diagnostic tests. For example, genomic and proteomic microarray technologies are expected to lead to diagnostic classification using multiple biomarkers fused by a computer that has been trained with previous cases with known truth status. This is the classical field of statistical learning machines. There is a formal correspondence between the problem of assessing these technologies and the MRMC ROC problem sketched above. Different radiologists are trained with different collections of patient cases—with differing amounts of total experience as well as range of case difficulty in their particular experience. The performance of a trained radiologist using a particular imaging technology is then assessed using new, i.e., previously unseen, cases. The correspondence with automated statistical learning machines is straightforward to make. The conclusion for our present purposes is that the problem of statistical learning machines is a member of the family of multiple-random-effects processes that we have encountered earlier with the MRMC ROC problem.

The most common form of assessment used for statistical learning machines is the method of cross-validation. In this method, the available samples (e.g., patient cases with known truth) are partitioned into, say, K groups. The algorithm is trained with $K - 1$ of the groups and tested on the remaining group; this is then repeated K times, with each of the groups taking a turn as testers. Cross-validation is a highly popular approach, especially in its extreme leave-one-out version. The well-known advantage of the latter approach is that its mean performance is characteristic of a training sample of almost the full size of the available sample (i.e., simply reduced by one sample) and, in that sense, it is almost unbiased.

There are several disadvantages of cross-validation that are not widely appreciated [32, 33]. The first is that it is not a smooth estimator: small changes in the data can lead to large differences in the results. This estimation method can therefore sometimes suffer much greater variability (e.g., larger variance and mean-square error [34]) than alternatives to be mentioned below. Moreover, one usually intends to use cross-validation to estimate the performance of a classifier *conditional on the available training set*. A very subtle and profound result is that cross-validation methods for regression and classification are only weakly correlated with this target conditional performance. The methods therefore hold no privileged position among the alternative methods of estimating performance. Finally, some investigators expect that the resampling involved in cross-validation incorporates a measure of the uncertainty over random training sets. This is patently incorrect since the different training sets used in cross-validation are almost the same and thus are far from random samples of trainers.

The points just listed were elaborated during the recent doctoral research of Waleed A. Yousef, Sc.D., a co-op student of Prof. Murray Loew of George Washington University and myself at CDRH/FDA. Details of this work are presented in several publications [35-38]. For our present purposes it will suffice to mention that Yousef has extended the approach of "generalized cross-validation" of Efron [32] and Efron and Tibshirani [33] to the problem of estimating ROC total and partial area in the classifier problem. In generalized cross-validation one draws from the available samples a bootstrap sample and uses this to train a classifier architecture of interest; the samples not drawn in the bootstrap sample are used as the testers—and the process is repeated many times. This is a smooth version of cross-validation that can be used in several different approaches to estimating the performance of a statistical learning machine. Since it is smooth, it can also be used as a starting point for estimating the statistical *influence function*—a variational/perturbation approach that allows one to estimate the *variance* of one's estimate of performance using differential calculus together with the original bootstrap samples.

The ability to estimate the uncertainty in one's estimate of performance provides the opportunity to use the results of a pilot study to size a larger pivotal study. One observes the variance in the pilot study and compares this to the variance or error bars one requires for the purposes of a pivotal trial. One can then estimate the size of the required trial compared to the original trial from the ratio of the variances—other things being equal.

In a pivotal trial one is usually expected to follow what we call the traditional hygiene, that is, the requirement of having completely independent training and testing sets. A formal approach to assessing performance in that setting—together with estimation of the resulting uncertainty—has also been presented by Yousef et al. [38]. This problem is formally very close to the MRMC paradigm discussed above.

Finally, Yousef et al. have also studied the partial area under the ROC curve for the classifier problem. Several surprising results have been reported [37]. Practical applications of the approach are being explored in the course of the current year.

12. THE TWO ELEPHANTS STILL IN THE LIVING ROOM

Our mentor Harold Kundel always emphasizes that in practice every gold standard (or "truthing" mechanism) is imperfect. He thus emphasizes that real-world studies are "agreement studies" – a concept toward which our community is ambivalent. One reason for this ambivalence is our understanding and experience that assessment in the absence of a good standard of truth will naturally suffer much greater uncertainty than assessment in the presence of known truth [39]. This uncertainty is equivalent to reduced statistical power for the comparison of interest in the study. The problem of imperfect truth is one of the elephants in the living room that has been brushed over lightly by our community.

Another elephant in the room has escaped our attention today because of time limitations, namely, the problem of location-specific scorekeeping and the associated issues of correlated (or "clustered") reports. Chakraborty has continued to make advances on this front and they can be reviewed by referring to his Website [40]. Chakraborty and Berbaum have recently presented validations of the JAFROC software under certain carefully prescribed scorekeeping rules [41]. They have acknowledged that the present approach does not yet address the problem of computer-aided detection where there are multiple false positives and one wishes to penalize an imaging system or computer aid for this. Many contemporary researchers are finding the JAFROC approach theoretically and practically useful for their investigations.

CONCLUSION

The field of multivariate ROC analysis and associated methods continues to offer solutions to current problems of interest in the fields of diagnostic testing and imaging, as well as related computerized detection and diagnostic tasks. It also continues to present us with an inviting horizon for continued investigations.

The PowerPoint slides for this presentation may be obtained by contacting the author: robert.wagner@fda.hhs.gov

REFERENCES

[1] Wagner R.F., Beiden S.V., Campbell G., Metz C.E., and Sacks W.M. "Assessment of medical imaging and computer-assist systems: Lessons from

[2] Wagner R.F., Beiden S.V., Campbell G., Metz C.E., and Sacks W.M. "Contemporary issues for experimental design in assessment of medical imaging and computer-assist systems." Proceedings of the SPIE Vol. 5034 (Medical Imaging 2003), p 213-224.

[3] Wagner R.F. "An Overview of Contemporary ROC Methodology in Medical Imaging and Computer-Assist Modalities." Powerpoint tutorial presented at the FDA Radiological Devices Public Panel Meeting of the Medical Devices Advisory Board, February 3, 2004. This can be accessed from the following link: http://www.fda.gov/ohrms/dockets/ac/04/slides/4024s1.htm

[4] Wagner RF, Metz CE, Campbell G. Assessment of Medical Imaging and Computer-Aids: A Tutorial Review. [In Preparation]

[5] Biggerstaff BJ. "Comparing diagnostic tests: a simple graphic using likelihood ratios." Statistics in Medicine 2000; 19: 649-663.

[6] Wagner RF, Beam CA, Beiden SV. "Reader Variability in Mammography and its Implications for Expected Utility over the Population of Readers and Cases." Medical Decision Making 2004; 24: 561-572.

[7] Metz CE. "Basic principles of ROC analysis." Seminars in Nuclear Medicine 1978; 8: 283-298.

[8] American College of Radiology. Breast imaging recording and data system (BIRADS). Reston, VA: American College of Radiology, 1993. (Cf. also latest edition, 2004.)

[9] Swets JA and Pickett RM. *Evaluation of Diagnostic Systems*. Academic Press, New York, 1982. (This ref. was developed as part of an historic project sponsored by the National Cancer Institute to develop a standard for evaluation of medical imaging. Details of this history are presented in the Preface and Acknowledgments of this book.)

[10] Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27: 723-731.

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

[12] Obuchowski NA and Rockette HE. "Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: An ANOVA approach with dependent observations." Communications in Statistics: Simulation and Computation 1995; 24: 285-308.

[13] Hillis SL, Obuchowski NA, Schartz KM, and Berbaum KS. "A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data." Statistics in Medicine 2005; 24: 1579-1607.

[14] Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Academic Radiology 2005; 12: 1534-1541.

[15] Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine (in press, 2006).

[16] Hillis SL, Berbaum KS. Recent developments in the Dorfman-Berbaum-Metz (DBM) procedure for multi-reader ROC study analysis. Presented at Medical Image Perception Society (MIPS) Conference XI, September 2005, Windermere UK.

[17] Barrett HH, Kupinski MA, Clarkson E. Probabilistic foundations of the MRMC method. Proc. of the SPIE 5749, Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment [Eckstein MP and Jiang Y, Eds] Bellingham WA 2005, pp21-31.

[18] Clarkson E, Kupinski MA, Barrett HH. A probabilistic development of the MRMC method. Statistics in Medicine [in review, 2005].

[19] Gallas BD. One-shot estimate of MRMC variance. Presented at Medical Image Perception Society (MIPS) Conference XI, September 2005, Windermere UK.

[20] Gallas BD. One-shot estimate of MRMC variance: AUC. Academic Radiology 2006; 13: 353-362.

[21] Campbell G, Douglas MA, and Bailey JJ. Nonparametric comparison of two tests of cardiac function on the same patients using the entire ROC curve. Proceedings of Computers in Cardiology, Washington DC, September 25-28, 1988: pp 267-270 (published by the IEEE Computer Society Press).

[22] Samuelson FW and Wagner RF. Bootstrapped MRMC confidence intervals. Proc. of the SPIE 5749, Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment [Eckstein MP and Jiang Y, Eds] Bellingham WA 2005, pp 15-20.

[23] Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects receiver operating characteristic analysis. Acad Radiol 2000; 7: 341-349.

[24] Bandos AI, Rockette HE, Gur D. Personal communication 2006.

[25] Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Suddhasatta A, Conant EF, Fajardo LL, Bassett L, D'Orsi C, Jong R, Rebner J, for the Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group. Diagnostic performance of digital versus film mammography for breast-cancer screening. N. Engl. J. Med. 2005; 353: 1773-1783. (Cf. also their published supplementary data.) Originally published at www.nejm.org September 16, 2005.

[26] Baker SG, Pinsky P. A proposed design and analysis for comparing digital and analog mammography: special ROC methods for cancer screening. J. Am. Stat. Assoc. 2001; 96: 421-428.

[27] Metz CE and Pesce L. Personal communication of expected availability of ROC software upgrades.

[28] Metz CE and Pan X. "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. J. Math. Psychol. 1999; 43: 1-33.

[29] Web address for University of Iowa software: ftp://perception.radiology.uiowa.edu/PUBLIC .

[30] Web address for University of Chicago ROC software:
http://www-radiology.uchicago.edu/krl/toppage11.htm#software.

[31] Wagner RF. Toward a strategy for consensus development on a quantitative approach to medical imaging. Academic Radiology 2006; 13: 137-139.

[32] Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. of the Amer. Stat. Assoc. 1983; 78: 316-331.

[33] Efron B and Tibshirani R. Improvements on cross-validation: The 0.632+ bootstrap method. J. of the Amer. Stat. Assoc. 1997; 92: 548-560.

[34] Sahiner B, Chan HP, Petrick N, Hadjiiski L, Paquerault S, and Gurcan MN. Resampling schemes for estimating the accuracy of a classifier designed with a limited data set. Medical Image Perception Society (MIPS) Conference XI, Airlie Conference Center, Warrenton VA September 2001.

[35] Yousef WA, Wagner RF, Loew MH. Comparison of non-parametric methods for assessing classifier performance in terms of ROC parameters. [In] *Applied Imagery Pattern Recognition, Proc. of the 33rd Workshop*; IEEE Computer Society, 2004, pp. 190-195.

[36] Yousef WA, Wagner RF, Loew MH. Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier. Pattern Recognition Letters 2005; 26: 2600-2610.

[37] Yousef WA, Wagner RF, Loew MH. The partial area under the ROC curve: Its properties and nonparametric estimation for assessing classifier performance. Submitted (2005) to the journal Pattern Recognition.

[38] Yousef WA, Wagner RF, Loew MH. Assessing classifiers from two independent data sets using ROC analysis: a nonparametric approach. Submitted (2005) to IEEE Trans. Patt. Anal. Machine Intelligence (PAMI).

[39] Beiden SV, Campbell G, Meier KL, and Wagner RF. On the problem of ROC analysis without truth: The EM algorithm and the information matrix. Proc. of the SPIE 2000; 3981: 126-134.

[40] http://www.devchakraborty.com This Website reviews the issues and provides regular up-dates on draft documents and the current status of the JAFROC software and related solutions.

[41] Chakraborty DP and Berbaum KS. Observer studies involving detection and localization: Modeling, analysis, and validation. Med. Phys. 2004; 31: 2313-2330.