Published in final edited form as:

Acad Radiol. 2007 July ; 14(7): 814–829. doi:10.1016/j.acra.2007.03.012.

Reliable and computationally efficient maximum-likelihood estimation of “proper” binormal ROC curves

Lorenzo L. Pesce, Ph.D. and Charles E. Metz, Ph.D.

Department of Radiology, MC 2026, The University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637-1470

Abstract

Rationale and Objectives—Estimation of ROC curves and their associated indices from experimental data can be problematic, especially in multi-reader, multi-case (MRMC) observer studies. Wilcoxon estimates of area under the curve (AUC) can be strongly biased with categorical data, whereas the conventional binormal ROC curve-fitting model may produce unrealistic fits. The “proper” binormal model (PBM) was introduced by Metz and Pan (1) to provide acceptable fits for both sturdy and problematic datasets, but other investigators found that its first software implementation was numerically unstable in some situations (2). Therefore, we created an entirely new algorithm to implement the PBM.

Materials and Methods—This paper describes in detail the new PBM curve-fitting algorithm, which was designed to perform successfully in all problematic situations encountered previously. Extensive testing was conducted also on a broad variety of simulated and real datasets. Windows, Linux, and Apple Macintosh OS X versions of the algorithm are available online at <http://xray.bsd.uchicago.edu/krl/>.

Results—Plots of fitted curves as well as summaries of AUC estimates and their standard errors are reported. The new algorithm never failed to converge and produced good fits for all of the several million datasets on which it was tested. For all but the most problematic datasets, the algorithm also produced very good estimates of AUC standard error. The AUC estimates compared well with Wilcoxon estimates for continuously -distributed data and are expected to be superior for categorical data.

Conclusion—This implementation of the PBM is reliable in a wide variety of ROC curve-fitting tasks.

Keywords

Receiver operating characteristic (ROC) analysis; receiver operating characteristic (ROC) curves; proper binormal model; maximum likelihood estimation (MLE); multi-reader; multi-case (MRMC) analysis

Corresponding author contact information: e-mail: E-mail: c-metz@uchicago.edu, voice: 773-702-6779, fax: 773-702-0371.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Radiological research often lies at the interface between human observers (radiologists) and technology. When technologies are applied to a population of patients, they usually produce extremely variable information (e.g., because both small and large lesions are represented), and in clinical settings this variable information is interpreted by human observers with different training, experience and capabilities. Such observers sometimes produce different final decisions when evaluating the same information; moreover, a single radiologist may report different results when reading a particular image on multiple occasions. Therefore, overall assessments of radiological technology often must deal with large variability, the modeling of which is complex (3), so multi-reader, multi-case (MRMC) methodologies are used increasingly to evaluate new diagnostic techniques (4–9).

MRMC and some simpler statistical analyses require estimation of a population figure of merit or index such as area under the ROC curve (AUC) for a subset of the data, conditional upon one or more covariates; in MRMC analysis one needs to estimate AUC, for example, for each imaging-modality/reader combination. Maximum-likelihood estimation (MLE) often is used to calculate these indices from a given dataset because of its good statistical properties and clear interpretation (10). MLE of ROC curves from categorical confidence-rating data was first employed by Dorfmann and Alf (10), who described a procedure based on the conventional binormal model (CvBM).

The CvBM uses two parameters to describe each ROC curve. The decision maker (e.g., radiologist) is assumed to sort cases into “positive” and “negative” groups by comparing the value of some internal decision variable for each case with a critical value, and by then deciding that a given case is “positive” if the value of its internal variable exceeds that critical value. The other basic assumption of the CvBM is that outcomes of the internal decision variable from actually negative and actually positive cases can be transformed by some single, unknown monotonic transformation to follow a pair of approximately normal (Gaussian) latent distributions, which are described jointly by the two parameters of the model. The values of these parameters of the transformed distributions and the model’s associated category - boundary parameters can then be determined from categorical confidence-rating data by MLE.

Empirical evidence indicates that two-parameter ROC curves describe a broad variety of observer data (11,12) with accuracy that can be considered sufficient for most practical purposes due to the substantial uncertainty that sampling variation causes in most real-world ROC studies. Although studies that involve thousands of cases may require use of more elaborate curve-fitting models, two-parameter models can be expected to provide adequate curve estimates for datasets that involve a few dozen to several hundred cases. An ROC-curve estimation algorithm also must be reliable when given small datasets and must perform consistently when given data collected on any of a variety of scales, such as ordinal-category or continuous, to allow imaging scientists and physicians to draw reliable conclusions. In particular, the CvBM and its extension to continuously-distributed data (13) have been applied successfully to a wide variety of problems in radiology (12,14,15), and until now they have served as the basis for the Dorfman/Berbaum/Metz (DBM) method for statistical analysis of MRMC data (4).

Despite its success, the CvBM suffers from the severe limitation that it can provide fitted ROC curves with highly unrealistic shapes in some situations. For mathematical reasons, essentially all CvBM ROC curves include at least a short portion that lies beneath the 45-degree guessing line and have a “hook” — i.e., a change from convex to concave curvature — at one end. Although for many datasets the magnitude of this effect is graphically invisible and has negligible impact on the interpretation of the empirical evidence, it can be problematic when

the number of cases is small and/or when the categories of a discrete ordinal scale are poorly distributed — i.e., when the distribution of empirical operating points along the ROC is very non-uniform. In such situations, the probability of estimating a curve with a large “hook” can be substantial even when the population ROC curve is convex (14), and the impact of this effect cannot be ignored if meaningful statistical inferences are to be drawn. All ideal-observer ROC curves are convex, and curves with hooks indicate that classification performance can be improved by intentionally introducing a random procedure into the decision process. Therefore, it is widely believed that forcing fitted ROC curves to be convex provides more reliable curve and accuracy-index values. Moreover, since different readers use ordinal category scales differently (e.g., different radiologists may disagree on whether particular mammograms should be classified in BI-RADS category 3 or 4), it can easily happen that one reader or one modality produces hooks consistently while another does not, thereby biasing any comparison of the associated ROC curve or index estimates.

In MRMC studies, the likelihood of encountering hooked ROC estimates with the conventional binormal model becomes considerable, in part because the number of cases used in such studies may be limited by logistical and economic constraints and in part because large numbers of readers are more likely to include substantially different spreads of empirical operating points. For example, the presence of hooks and occasional failure of the DBM algorithm to produce MLEs for some reader-modality combinations may explain the unusually large variance of MLE indices as opposed to nonparametric Wilcoxon estimates that has been reported in one MRMC comparison study (16).

To address these issues, Metz and Pan (1) developed a “proper” binormal model (PBM) and software entitled PROPROC that used this model to fit convex ROC curves by MLE. Although the PBM is based upon use of likelihood ratio as the decision variable, and as such is an ideal-observer model, in practice this model does not assume that the human decision maker is an ideal observer, but rather that the human observer’s ROC curve *could* have been produced by an ideal observer — e.g., if given less information than was available to the human observer. This assumption does not seem strong, and evidence supports its promised flexibility (15). In effect, the PBM assumes that human observers base their decisions primarily upon a perceived likelihood of abnormality rather than directly upon a quantity such that both very small and very large values, for example, constitute strong evidence of abnormality. The PBM is called “proper” because it forces the curve shape always to be convex. Accordingly a proper ROC curve cannot have hooks, drop below the 45° chance line or have an AUC smaller than 0.5. The latter situation can be modeled as misused detection information, which makes sense because to make a consistently wrong choice one needs to be able to distinguish consistently between the two states of truth. Study of the paper by Metz and Pan (1) is encouraged as a basis for full appreciation of the work reported here.

Although the implementation of the PBM described by Metz and Pan (1) performed well for simulated datasets and many real datasets (17), subsequent applications of that software indicated that numerical and programming-related issues rendered its use problematic in some extreme situations (2). Moreover, that early software was unable to provide reliable of standard error by inversion of a Hessian matrix, as described by Dorfman and Alf for the CvBM (10). Therefore, we redesigned the PBM curve-fitting algorithm almost completely and tested it extensively on a very broad variety of simulated and real datasets. On the basis of those studies, we believe that the PBM now satisfies the reliability requirements of MRMC analysis. Consequently, the PBM is now included as a curve-fitting module in the DBM-MRMC2.0 software package that we have developed in collaboration with Professor Kevin Berbaum’s group at the University of Iowa, which uses this version of the PBM and can be downloaded at <http://xray.bsd.uchicago.edu/krl/>. Our modifications of the original algorithm and the results of our simulation studies are reported below.

Materials and Methods

All modifications of the original algorithm described here were necessary either to minimize bias in the estimates of AUC and of their standard errors or to render the computation numerically more stable, and many of the modifications served both purposes. Therefore, we take a structural approach in the following, rather than try to rank the changes in terms of importance. The code was rewritten strictly according to the standards of FORTRAN f95, a modern, reliable and highly portable language that is well suited to high-precision numerical calculations. We believe that a detailed description of our implementation of this MLE ROC curve-fitting algorithm is important due to our conviction that its strength has potentially broad applications in ROC analysis and can be used to render also other MLE curve-fitting algorithms more stable. For clarity, the description of our methods in this section occasionally mentions quantitative measures of the new algorithm's performance. A more complete report of our evaluation of the algorithm is provided in the subsequent Results section, however.

Data categorization scheme

When ROC data are comprised of continuously or quasi-continuously distributed values, the LABROC4 categorization scheme (13) can be used to convert them into ordinal-category data without loss of information that is relevant to ROC curve estimation. This scheme first retains only the corners of the empirical ROC that results when operating points are plotted directly from the raw data at all possible critical-value settings, and then interprets the horizontal and vertical coordinates of those corner points (weighted by the numbers of actually negative and actually positive cases, respectively) as rating data that were collected on a discrete ordinal category scale. The same method can be applied also to collapse uninformative categories in inherently categorical data — e.g., unused categories or adjacent categories that contain responses only from the same truth state.¹ The resulting categorical data are then fit by MLE. However, in contrast to the LABROC4 curve-fitting algorithm described by Metz et al. (13), we employed the PBM as parameterized by Metz and Pan (1).

The original PROPROC algorithm described in (1) employed a second categorization step that was intended to improve numerical stability: after categorical data had been obtained according to the LABROC4 scheme, they were collapsed again into a smaller number of categories (at most 10 or 20) chosen upon the basis of the number of cases in the dataset and the total number of categories produced by the LABROC4 scheme. However, exploratory calculations with real datasets showed that this additional collapsing can alter the value of the final AUC estimates by as much as several percent when the number of LABROC4 categories is on the order to 15 to 50 and can alter curve shape considerably, usually through elimination of one or a few empirical operating points that would otherwise constrain the fit.

Computer simulations that sampled large numbers of datasets from a known population showed that use of additional category reduction schemes like that described immediately above biases the estimates in a complicated, almost unpredictable, way. In contrast, our new algorithm for fitting proper binormal ROC curves is sufficiently stable to allow full use of the LABROC4-categorized data for up to 400 categories, as in the LABROC4 algorithm that employed the CvBM (13). Extensive simulation studies with the new PBM algorithm indicate that when a LABROC4-categorized dataset involves up to 400 categories, the order of the resulting

¹One must note that the information discussed here concerns only ROC curve estimation in binary classification tasks, rather than assessments of clinical effectiveness, in which maintaining distinctions between adjacent categories of response may be important. Indices of diagnostic performance that merge detection and effectiveness considerations also may use the information in adjacent response categories directly (18). Other categorization schemes are likely to be more appropriate also when truth is not considered to be binary; some approaches that address this situation are similar to the one just described, mixing detection and effectiveness considerations (19), whereas others attempt to generalize ROC analysis by taking an explicit multi-class approach to truth (20).

estimation bias is never greater than that of the error obtained after optimization of the algorithm's double-precision floating-point numerical representation, which is accurate to 6 or 7 decimal places for AUC and parameter estimates. From a practical perspective, such biases or numerical errors correspond to undetectable differences in fit, because many millions of cases would be needed to yield a fit with a standard error of this magnitude. Moreover, no inevitably idealized mathematical model should be depended upon to provide reliable fits to huge real-world datasets.

Computational speed must be considered as well. Our new PBM algorithm involves computation and inversion of the Hessian matrix of second-order partial derivatives of the log-likelihood function with respect to the model parameters. With an n -category dataset, the elements of this real symmetric matrix are represented by a total of $n(n+1)/2$ real numbers, many of which must be computed as a sum of $2n$ terms. Calculation and inversion of the Hessian matrix scale approximately as the third power of n , so a 400-category dataset requires approximately 10,000 times as much CPU time as a 20-category dataset. After refinement of our new PBM algorithm, we found this to be an issue only in large simulation studies, because estimation of a single ROC curve from even very large, continuously-distributed and numerically challenging datasets never required more than a minute on an ordinary Pentium 4 desktop computer. Real MRMC datasets usually involve no more than 10 to 30 employed categories and therefore can be fit in several milliseconds. Nevertheless, our new algorithm allows the user to constrain the maximum number of categories in the data to be analyzed, thereby permitting faster fits after due warnings concerning the possibility of increased bias and/or reliability.

In a large number of large simulation studies involving continuous data generated from the PBM and more than 1000 cases, many of which led to more than 350 LABROC4-scheme categories, we never found any numerical instability below 400 categories. In fact the algorithm was rarely unstable even with more than 400 categories, but we decided to set the maximum allowed number of categories to 400, a value that never produced instability. One should note that any ROC curve-fitting calculations by MLE can be unstable when very large datasets arise from a model different from that used to fit the data, thereby causing at least some datasets to be highly unlikely given the assumed model. However, our new PBM algorithm has been tested successfully on several real datasets involving thousands of cases without detectable problems.

Numerical issues with functions related to the proper binormal model

The PBM curve-fitting computations involve evaluation of a number of special functions and differences among them (1). Our new algorithm's subroutines related to normal (Gaussian) distribution functions were obtained from the Anderson Cancer Center (<http://biostatistics.mdanderson.org/SoftwareDownload/>). Care was taken to employ asymptotic forms when the dependent value of a computed function approached numerical or mathematical limits — e.g., in evaluation of the cumulative univariate standard-normal distribution function when the result was close to 0 or 1, and in calculating differences between those cumulative distributions when the differences were much smaller than individual distribution-function values. Particular attention was paid also to evaluation of the cumulative *bivariate* standard-normal distribution function, which is used to compute AUC for proper binormal ROC curves (1), to guard against numerical imprecision when the correlation coefficient input value is very close to 1.

Most of these issues turned out to be more important in unveiling problems in the old code than in developing the new algorithm. For example, it was originally believed (1) that the (d_a, c) parametrization of the PBM and the two-step iterations described by Metz and Pan (1) were necessary to render the calculations stable enough to allow iteration to converge. However, during the work reported here we discovered that instabilities in the original code

occurred primarily in obtaining initial estimates of the category-boundary parameters, rather than during subsequent iteration. A different approach to the task of obtaining initial parameter estimates, described next, caused many other problems to disappear.

Initial estimates

In general, MLE of ROC curves (like virtually all other iterative optimization procedures) requires initial estimates: parameter values from which an iterative scheme begins to search for the maximum of the log-likelihood function. In the case of ROC-curve models that involve latent data categories, an iterative scheme needs initial estimates both for the curve parameters (d_a and c in the proper binormal model) and for the “cutoffs” — i.e., the boundaries between decision categories (1,10).

With the CvBM, each ROC curve is specified by a pair of latent normal (Gaussian) distributions, which are characterized jointly by two curve parameters (often denoted by a and b), and by a cutoff parameter that specifies any single combination of false positive fraction (FPF) and true positive fraction (TPF) on that curve. All population ROC curves that arise from the binormal model plot as straight lines on normal-deviate axes (21), so a straight line can be fit to empirical operating points plotted on such axes by the ordinary method of least squares to obtain initial estimates for subsequent fitting of CvBM ROC curves by MLE (10). These fits are usually quite reasonable and serve well as a basis for initial curve and cutoff estimates. Unfortunately, no analogous procedure for least-squares fitting of a straight line is known for the PBM, so some less direct approach must be used to obtain initial estimates of this model's ROC curve and cutoff parameters.

We discarded the informed grid-search technique described in (1) because of its complexity and inefficiency. Moreover, our simulation studies indicated that the grid search sometimes was unable to identify curve parameter values in the neighborhood of the MLE reliably because the log-likelihood function of the proper model can have long, narrow and twisted ridges (1), which are typically difficult to track on a grid unless the latter is very finely meshed and, hence, computationally expensive. Worse still, we found that a grid search often becomes numerically unstable because the log-likelihood function must be evaluated even at highly unlikely combinations of parameter values.

Perhaps surprisingly, we found that fitting a straight line on normal-deviate axes to a dataset's empirical operating points by ordinary least squares — i.e., obtaining initial ROC curve parameter estimates for the CvBM rather than for the PBM — served well also in choosing initial curve-parameter estimates for the PBM. Because of known relationships between the curve parameters of the conventional and proper binormal models (1), a simple least-squares fit provides a good indication of whether the CvBM and PBM will yield similar or substantially different MLE fits and, thereby, provides a rough approximation of the PBM curve parameters' MLE values.²

During our extensive simulation studies with the new algorithm, we noticed that the true log-likelihood function often can have up to four local maxima when datasets are small and highly unbalanced (i.e., the proportion of actually positive cases is far from 50%). However, a clear global maximum could be identified automatically for each of the simulated datasets by using multiple initial estimates, as described below. Usually the other maxima were located far from

²Metz and Pan (1) have shown that hooks are evident in a plotted CvBM ROC curve if and only if the parameter values of the corresponding PBM curve satisfy $|c| > 2d_a$. The combination of d_a and c associated with a least-squares CvBM fit does not provide reliable initial curve-parameter estimates for PBM MLE in such situations. The parameter d_a then has less influence than c on the PBM's ROC curves, however, so we set d_a equal to zero when $|c| > 2d_a$ according to the least-squares CvBM fit. A new least-squares estimate of c can then be calculated easily from the simplified relationship between FPF and TPF on a PBM ROC curve, thereby providing an empirically reliable initial estimate of the PBM ROC curve.

the global-maximum solution and provided highly skewed ROC curves or curves skewed in the wrong direction, which yielded local maxima in the log-likelihood function due to geometrical constraints. Occasionally but rarely, we found one or more lower satellite maxima near the global maximum. However, the latter situation occurred only when the case sample was small and highly unbalanced (e.g., 100 negative and 10 positive cases) and sequences of empirical operating points happened by chance to align on PBM ROC curves that are distinct but have similar skew.

We addressed the problem of multiple maxima in the likelihood function as follows. Because of geometrical constraints in the ROC space, the fits associated with these multiple maxima tend to have similar values of AUC. Therefore, the corresponding values of d_a and c to lie near a single isopleth of AUC on a plot of c versus d_a , with each isopleth resembling a flattened ellipse (1). By choosing three initial pairs of PBM curve-parameter estimates, one that corresponds to a symmetrical ROC curve ($c = 0$) and two others skewed in opposite directions ($c > 0$ and $c < 0$) with the same AUC, an iterative optimization scheme is able to converge upon all possible locations of maxima if well-separated multiple maxima exist and, thus, to choose the global maximum from among them. Well-separated multiple maxima are sometimes accompanied by additional “satellite” maxima located nearby. Therefore, when multiple maxima are identified by the procedure just described, a set of 5 additional initial estimates is distributed along an elliptical arc that approximates an isopleth of AUC to determine whether satellite maxima are present and, if so, which is the true global maximum. This search algorithm was tested thoroughly and found to locate the global maximum consistently when multiple maxima are present.

After initial estimates of the PBM ROC curve parameters have been obtained, the algorithm proceeds to compute initial estimates for the cutoffs. The latter parameters are initialized individually in essentially the same way as that used previously in the LABROC4 algorithm (13). For each cutoff parameter, the categorized data are further collapsed into only two categories that lie on either side of the cutoff in question. Given initial estimates of the PBM ROC curve parameters from the least-squares procedure described above, an MLE estimate of the single cutoff parameter is then computed by solving numerically the nonlinear equation that results from differentiating the expression for log-likelihood with respect to the single cutoff parameter in question and setting the result equal to zero. We found this approach to be simple, robust and precise enough to provide consistently good initial estimates of the cutoff parameter values.

Proper binormal ROC curves are skewed with respect to the negative diagonal of the unit square when the curve parameter c is non-zero, with the magnitude and sign of c indicating the amount and direction of skew, respectively. One can show easily that changing the sign of c reflects a PBM ROC curve across the negative diagonal mentioned previously, which is equivalent to replacing the FPF and TPF values of each ROC point by $1 - \text{TPF}$ and $1 - \text{FPF}$, respectively. Unlike those of the CvBM, the cutoff parameter values of the PBM do not range from negative infinity to positive infinity, but instead have a finite maximum possible value (when $c > 0$) or minimum value (when $c < 0$). We exploited the symmetry associated with the parameter c to create a stable scheme for computation of initial cutoff estimates that ensures symmetrical results when empirical operating-point data are reflected across the negative diagonal of the unit square. Initialization begins from the finite bound (maximum or minimum), which is determined from the previously-obtained initial estimates of the PBM ROC curve parameters, d_a and c . This is considered the upper or lower bound for the first cutoff estimate, depending upon the sign of c . The maximum of each two-category log-likelihood function is then bracketed by moving away from the bound until log-likelihood decreases. The two-category MLE of the single cutoff parameter is then estimated by the Brent method due to its robustness

and efficiency (22) and used subsequently as a bound for the next cutoff, with this procedure continued until initial values have been obtained for all of the cutoff parameters.

Optimization

As mentioned above, we found the two-step iteration procedure employed in the original PROPROC algorithm (1) to be unnecessary after our new procedure for obtaining initial parameter estimates had been implemented. Therefore, we replaced that optimization scheme with the single-step Newton-Raphson method, which iterates all PBM parameter values simultaneously by using the inverse of the Hessian matrix of expected second-order derivatives of the log-likelihood function (i.e., Fisher scoring). More specifically, the new algorithm employs the Transaction on Mathematical Software algorithm TOMS 611, an implementation of a trust zone Newton-Raphson procedure (23). Our simulations and previous literature (23) indicate that this is a very stable procedure in many kinds of optimization tasks.

Several issues had to be addressed to render this efficient optimization possible. First, as already mentioned, care was taken in producing reliable initial estimates for the cutoffs. Second, since Newton-Raphson procedures usually find the local maximum closest to the initial parameter estimates, selection of a non-global maximum must be guarded against when multiple maxima are present. Our approach to the latter concern was described in the previous section, and our simulation studies indicate that it has been successful. Third, mathematically impossible values for the parameters had to be avoided, because our Newton-Raphson algorithm chooses an optimization step according to trust zone conditions and the product of the Hessian-matrix inverse and the local gradient vector, rather than according to the meaningfulness of the log-likelihood function if the proposed step were taken. Hence, our algorithm accepts the optimizer's next step only if the proposed $d_a > 0$, $|c| < 1$, and all proposed cutoffs would satisfy the range condition that the proposed curve-parameter values impose. Fourth, the cutoffs must be prevented from crossing. Occasionally, because of the local curvature of the log-likelihood function, a proposed step in one cutoff is sufficiently large to make it become smaller or larger than the cutoff below or above it (e.g., the proposed cutoff between category 3 and 4 may have a value larger than that of the cutoff between category 4 and 5). This is mathematically meaningless because it associates a negative probability with one or more categories. A more subtle but related requirement is to prevent cutoffs from becoming too close, which can cause numerical instabilities if the corresponding FPF and TPF of adjacent estimated expected operating points share the numerical values, given their computational precision. Fifth, occasionally the Newton-Raphson optimization procedure converges to a false maximum (detected automatically by the TOMS 611 algorithm), usually due to peculiar datasets that produce numerical instabilities in specific regions of parameter space. Restarting the optimization (occasionally more than once) with slightly altered parameter values always solved this problem.

We used the stopping criteria of the TOMS 611 optimizer (23) directly, because they were developed for similar problems and are both theoretically and empirically effective in locating maxima of log-likelihood functions. The resulting PBM curve-fitting algorithm is about 40 times faster than the previous scheme on average, and occasionally 200 times faster. Although optimization becomes much slower when multiple maxima are present, our simulations showed that the old algorithm could — and often did — miss the global maximum entirely for especially difficult datasets.

One should note that occasionally the MLE solution involves $d_a = 0$, which corresponds to the most highly-skewed ROC curve that the PBM can produce for a given AUC (1).³ When the d_a parameter is sufficiently small, it becomes essentially redundant in the sense that changing its value appreciably does not affect the ROC curve shape to any practical extent. This means that some distinct combinations of ROC curve-parameter values produce very nearly the same

curve; hence, the Hessian matrix becomes numerically singular, corresponding to the fact that curvature of the log-likelihood in one direction is very nearly equal to zero. The optimization procedure will halt in such situations, often failing to converge to the theoretical MLE solution. However, because differences between the computed and true maximum-likelihood ROC *curve* estimates are vanishingly small in such situations even though the parameter values may differ appreciably, the iterative approach described here consistently produces numerically correct MLEs of ROC *curves* and their summary indices, such as AUC.

Standard-error estimation

In order to draw meaningful statistical inferences, standard errors or variances must be estimated for the index of accuracy (e.g., AUC) that is to be tested. One should note that the DBM method does not require variance to be estimated during curve fitting because it is estimated instead by jackknifing (4). Nevertheless, there are many other situations where estimation of the standard error of a single AUC estimate, for example, is either necessary or useful.

To estimate the standard error of any index estimate without using a non-parametric method (24), which can be unreliable for inherently categorical data, or computationally intensive resampling (25), the variances of estimates of the parameters of the ROC curve must be determined first, because derived indices are functions of these parameters. Variances of the curve-parameter estimates are estimated from the inverse of the Hessian matrix after iteration stops, which yields the Cramer-Rao bound (26) on those variances. This is a lower bound, so when the curve-parameter variance is large (e.g., due to a small dataset), non-quadratic behavior of the log-likelihood function has a greater effect, and variance departs more strongly from the bound. Using the delta method (27), which is essentially a second-order Taylor-series expansion of an index value in terms of the ROC curve parameters, one can then obtain a reasonable estimate of index variance in most situations. The delta method and the estimates of the parameter variances themselves become less reliable when local behavior of the log-likelihood function departs substantially from quadratic, because the second-order expansion then becomes less accurate. Still, our simulation studies showed that the estimated variances in AUC were within 2 to 5% of their population variance in most situations. Estimated variances of the ROC curve parameters themselves were less reliable, however.

Estimates of standard error could not be provided reliably by the previous version of our proper-binormal ROC curve fitting software (1) because of the need to invert the Hessian matrix, which becomes very nearly singular when the maximum of the log-likelihood function lies in a region of parameter space where the ROC curve changes very slowly with d_a . The variance in d_a estimated by inverting the Hessian is nearly infinite in such situations because the log-likelihood function is nearly flat near its MLE in that direction. However, because this direction is redundant, its variance in fact has virtually no effect on estimation of the ROC *curve* for the dataset in question, and consequently has no appreciable effect on the variance of its derived indices. Therefore, our new algorithm eliminates the contribution of this direction to the inverse of the Hessian matrix, which is achieved by first computing the eigenvalues of the matrix and then eliminating the eigenvector corresponding to the smallest eigenvalue from calculation of the inverse if that eigenvalue is less than 10^{-6} times the value of the next-larger one — i.e., by calculating the Moore-Penrose pseudo-inverse (28,29) rather than the true inverse of the Hessian matrix. Our simulation studies have shown that this approach works very well in providing reliable estimates of the standard error of AUC for all of the population ROC curves tested. These results indicate that relying upon PBM curve parameters to compare datasets may

³The height of a proper-binormal ROC curve, unlike that of a conventional ROC, is determined by both d_a and c , whereas the curve's skew is determined only by c . Hence, $d_a = 0$ does not yield the chance-line ROC unless $c = 0$ as well.

be ineffective, and that a summary index, such as AUC or partial area (30,31), should be used instead. We note in passing that comparing AUC values when estimated ROC curves have very different shapes may not rank imaging modalities meaningfully because the curves are then likely to cross (31).

Evaluation studies

The new algorithm's performance was assessed by using it to analyze both empirical and computer-simulated data. We tested the algorithm on several especially stressful empirical datasets that the previous PBM curve-fitting algorithm (1) had found problematic (2) and on a variety of other real datasets from automated classifiers and human observers. However, only simulation studies allow one to know whether an algorithm "gets the right answer" in a statistical sense, because the population values are then known. Emphasis in these simulations was placed also on assessing the algorithm's reliability when confronted with "degenerate" datasets (14), poorly spread categorical data, and other numerical challenges. Such testing is especially important for a curve-fitting algorithm that is to be used in MRMC analysis. Simulations were performed for population PBM ROC curves of various heights and skewed in both directions ($c < 0$ and $c > 0$). The details of our simulation scheme are described next.

The PBM was used to generate simulated data for testing the performance of the new algorithm in terms of convergence, bias, standard error, and root mean square error (RMSE). Additional data-generating models would be needed in order to assess the robustness of the PBM in situations where other ROC curve-fitting models are, in fact, more appropriate. However, the results of our evaluation of the PBM with real datasets, reported below, indicate that the PBM is adequate in a broad variety of situations, especially for datasets of the sizes typical in assessments of radiological imaging. More detailed assessment of cross-model robustness is beyond the scope of this paper and must await further investigation. Accordingly, all of the simulation studies reported here assume, in effect, that a real dataset's values (e.g., confidence ratings) are produced from some monotonic transformation of a binormal random variable by calculating likelihood ratio. The latter step constitutes the key distinction between the CvBM and the PBM.

Large numbers of datasets were generated to mimic the experimental sampling of data from a real population. Moreover, datasets with different numbers of actually positive and actually negative cases were generated to simulate different experimental conditions. These datasets were generated from a broad range of different population ROC curves — with a variety of AUCs and either unskewed or skewed to a variety of extents — to simulate different kind of medical devices or diagnostic tasks and radiologists with different levels of expertise.

Each dataset was constructed by first generating two sets of outcomes of normally-distributed random variables using a random number generator from CPAN (<http://search.cpan.org/~grommel/Math-Random-0.67/Random.pm>). The specified properties of this random number generator appeared to be suitable for our research (see the CPAN website for details), and analyses of results from our simulations not reported here support the choice. The two sets of normal random variables for any population PBM ROC were generated from a CvBM with parameters a and b — i.e., the actually negative cases arose from a distribution with mean 0 and standard deviation 1, whereas the actually positive cases are sampled from a distribution with mean a/b and standard deviation $1/b$. The parameters of the corresponding PBM population ROC curve (1) are then given by $c = (b - 1)/(b + 1)$ and $d_a = av2/v(1 + b^2)$, whereas the PBM population AUC is given by Eqn. 36 in (1). The results reported below focus on the accuracy and precision with which AUC can be estimated for various PBM ROC curves, each specified by a combination of c and d_a .

Each outcome of a CvBM random variable is then transformed into a monotonic function of log-likelihood ratio according to Eqn. 11 in (1): $y = -[(bx - a)^2 - x^2]/2$, where x represents an outcome of the CvBM normal random variable and y is the corresponding PBM datum that serves as input to the curve-fitting algorithm. The latter step assures that the resulting population ROC curve in fact arises from the PBM. One should note that this scheme cannot be used to generate curves with $AUC = .5$ because y would be consistently 0. Because the CvBM ROC curve with $AUC = .5$ is already proper, it is sufficient to use the x values directly as data if PBM curves are to be generated with $AUC = .5$.

ROC curve fits obtained from these simulated data were summarized in terms of averages of AUC estimates, empirical standard deviations of those estimates, averages of the Hermitian standard-error estimates obtained from individual curve fits, and ratios of the empirical standard deviation of AUC estimates to the empirical standard deviation of Wilcoxon AUC estimates. In addition, we used pairs of independent AUC estimates from a common population ROC and their associated Hermitian standard-error estimates to perform a two-tailed statistical significance test of the difference in AUC and then plotted the relationship between empirical Type I error rate and α (i.e., the critical p-value), as described in (32). This statistical test was performed for each pair of PBM ROC curve estimates by evaluating the test statistic $z = [AUC_1 - AUC_2]/\sqrt{(SE_1)^2 + (SE_2)^2}$ and then using the standard-normal distribution function to determine whether the difference was statistically significant at the $100(1-\alpha)\%$ confidence level according to a two-tailed z-score test. Because the null hypothesis of no difference in AUC is in fact true for each such comparison, Type I error rate will equal α (within binomial variation that depends upon the number of simulated pairs) if the test statistic is normally distributed and the standard error of AUC is estimated accurately for each dataset.

Results

The major concerns expressed in the past regarding practical use of the proper binormal model have involved its ability to fit real data, lack of robustness and lack of standard-error estimates, so we emphasize those issues here. All of the results reported here were obtained with f95 FORTRAN code running on a Linux platform. However, we have compiled other versions of our PBM code for Windows (tested on Windows 2000 and later) and Apple OS X (tested on OS 10.4), and they produce the same results to within five decimal places for all of the datasets tested.

Real data

Some of the human-observer data (2) that caused problems for the previous PBM implementation are shown in Fig. 1. The new algorithm performed well on all problematic datasets identified in that study: each of the PBM fits in Fig. 1 passes close to the empirical operating points, lies well within their rather large error bars (not indicated for clarity) and provides ROC curves of reasonable shape, whereas the CvBM fits and the Wilcoxon interpolations clearly underestimate AUC. In addition, the new code provided a standard-error estimate for each fit that we considered reasonable. These datasets contain little information about the population from which they were sampled: the PBM's log-likelihood function is extremely flat (and there is no reason to expect the likelihood of other models to be less flat), so a variety of quite different curves fit the data nearly as well. This puts a large uncertainty on any curve estimate and makes any inference of difference extremely difficult. The proper-binormal parameter values estimated from the datasets shown in the first row of Fig. 1 were used also to specify population ROC curves for a simulation study. For each of those problematic situations, 5 groups of 10,000 datasets were generated randomly, as discussed in the previous section, by using the fitted PBM ROC curves as population curves with different

numbers of actually positive and actually negative cases in each group. No fitting problem was observed for any of those simulated datasets.

We did not conduct extensive simulation studies to evaluate systematically the robustness of the new algorithm when it is presented with data that did not arise from the PBM, due to the difficulty of choosing a meaningful population model from which to simulate data for that task and because no reliable goodness-of-fit measure has been developed for ROC curves. Instead, we took a more anecdotal approach by using the new algorithm to fit several dozen empirical datasets produced by human observers or automated classifiers. Figure 2 displays the results that we obtained for what might be consider the most problematic such data, reported by Kupinski et. al (33) from a study on an automated classifier. This dataset contained 647 actually negative (non-malignant) and 77 actually positive (malignant) cases, and so might be considered large. However, the limited number of actually positive cases yields the substantial ± 2 SE binomial error bars for each empirical operating point shown in Fig. 2, which were calculated from the expressions (34,35) $SE\{FPP_i\} = \sqrt{FPP_i(1-FPP_i)/(N_{NEG}-1)}$ and $SE\{TPF_i\} = \sqrt{TPF_i(1-TPF_i)/(N_{POS}-1)}$, where FPP_i and TPF_i are the coordinates of a particular point. One should note that error bars obtained in this way tend to be conservative, because the calculations assume independent classifier reports, whereas multiple reports in a single image may be correlated to some degree. Still, as can be seen in Fig. 2, the fitted curve lies within the error bars calculated for each point. When we generated simulated datasets of identical size using the MLE PBM curve in Fig. 2 as the population ROC, excursions from the fitted curve much like those seen in Fig. 2 often appeared in the resulting empirical ROCs, indicating that the PBM is in fact an adequate model for those data, given their statistical precision.

Comparing AUC estimates obtained with a non-parametric approach to those from our semi-parametric curve-fitting technique provides a way of assessing how far from the empirical points the fitted curve lies on average. The Wilcoxon (trapezoidal) and PBM estimates of AUC obtained from the data of Fig. 2 are 0.869 and 0.873, respectively, with estimated standard errors of .026 (according to the method described in (24)) and .013, respectively. Many additional datasets from Computer Aided Diagnosis (CAD) studies conducted in our laboratory by other investigators have been tested, and all produced fits similar to or better than that shown in Fig. 2.

Because one of our main goals in developing a reliable PBM curve-fitting algorithm was to allow its use in MRMC analysis, we evaluated the algorithm also with datasets obtained from other radiological human-observer studies. A variety of fully anonymized datasets were generously provided to us for this purpose by colleagues at The University of Chicago and elsewhere. Data collected on both discrete and quasi-continuous scales were included. The small to moderately-sized datasets of this kind that we studied always produced acceptable fits with the PBM, both because the fitting model is reasonable and because the error bars associated with the empirical points were substantial (11). For example, Fig. 3 displays our results for the two datasets collected on a quasi-continuous scale by Shiraishi et al. (36) that showed the greatest subjective departures from our algorithm's fit. The PBM fit for the data of Reader 7, in the left panel, lies within most of the binomial $\pm 2SE$ error bars. Although better fits usually resulted when simulated datasets of the same size were generated from the PBM by using the real-data MLE fit as the population ROC, the discrepancy seen in the left panel of Fig. 3 is not surprising because that fit was the worst in a group of 20. More interesting, perhaps, are the data from Reader 13, shown in the right panel of Fig. 3, where the fit is still reasonable but some empirical operating points lie on the $FPP = 0$ or $TPF = 1$ axis and the fitted PBM ROC lies slightly below the error bars near $TPF = 1$. However, this discrepancy is tempered by the observation that binomial $\pm 2SE$ error bars do not approximate 95% confidence intervals well when empirical TPF or FPP is near 0 or 1. More quantitative assessments of such

discrepancies must await the development of meaningful goodness-of-fit measures in ROC analysis.

Continuously-distributed simulation data

Our most systematic simulation studies investigated three AUC values to cover the range of practical interest in medical imaging: .6 (low), .85 (typical), and .95 (high). To simulate a range of curve shapes as well, we included population ROC curves with different amounts of skew: symmetrical, moderately skewed, and highly skewed. [No standard exists for quantifying the skew of an ROC curve, so we use the term here in a subjective way.] Figure 4 shows the resulting variety of population ROC curves from which we simulated continuously-distributed data, with the corresponding curve parameter values listed in Table 1. Our studies indicated clearly that the new algorithm's results are perfectly symmetrical when the sign of the PBM curve parameter c is changed, so results are reported here only for $c < 0$.

As described in (1), substantial changes in the PBM curve parameter d_a have very little effect on the ROC curve if $|c| > 6d_a$. Therefore, large uncertainty in PBM ROC curve-parameter estimates is not necessarily problematic; a statistically significant difference between the parameter values of two PBM ROC curves does not necessarily correspond to a statistically significant difference between the curves themselves; and estimated PBM curve parameter values must be used with great caution, if at all, to draw inferences concerning population ROC curve location or shape. At least until this issue is better understood, we recommend that investigators test the statistical significance of differences between PBM curve estimates not in terms of their curve parameter values, but instead in terms of AUC if a global index is considered adequate after thoughtful consideration, or otherwise in terms of a more localized index such as partial area that can be linked to the needs of a particular clinical task. Therefore, the results reported here focus on AUC.

Tables 2–4 report concisely the results of our systematic simulation studies in terms of averages of AUC estimates, empirical standard deviations of those estimates, averages of the Hermitian standard-error estimates obtained from individual curve fits, and ratios of the empirical standard deviation of AUC estimates to the empirical standard deviation of Wilcoxon AUC estimates.⁴ As one can demonstrate easily from the data in these tables, bias was very nearly proportional to $1/N$, where N represents the total number of cases (half of which were actually positive in each dataset), whereas standard error varied very nearly as $1/\sqrt{N}$. Accordingly, bias becomes less and less important relative to standard error as dataset size increases. Bias and standard error in AUC estimates obtained from highly unbalanced datasets require further investigation, but our preliminary results indicate that the proportional relationships described above hold over a substantial range of prevalences if the fraction of actually positive cases in a dataset is held constant as dataset size is varied.

As one might expect when a relevant parametric or semi-parametric model is employed, Tables 2–4 show that the standard errors of the MLE estimates of AUC tended to be smaller than those of the non-parametric Wilcoxon estimates. An observation that the strongest bias occurred when $AUC = .6$ is explained by the fact that estimated AUC has a lower bound of .5 with the PBM, thereby reducing variance but increasing bias when sampling variation in small datasets caused empirical operating points to fall substantially below the positive diagonal of the unit square. Indeed, bias in the PBM MLE estimates was not small with few cases and skewed population ROC curves. This is probably caused by a combination of the lower bound on AUC estimates ($\geq .5$ for any proper model) and the bounds on curve shape (which is most strongly

⁴Averages of the Wilcoxon AUC estimates are not reported here because that estimator is known theoretically (37), and was shown empirically in our studies, to be unbiased for continuously-distributed data.

skewed with the PBM when $d_a = 0$). In contrast, the non-parametric Wilcoxon (trapezoidal-area) estimator is completely unbiased for continuously-distributed data (37) but can be very strongly biased toward low AUC estimates when inherently categorical data are collected in a small number of categories, as one can see easily from the empirical operating points shown in Fig. 1. Although detailed comparisons of the PBM and Wilcoxon approaches are still underway, we conclude from the results presented here that the performance of our new PBM MLE algorithm is roughly equivalent to that of the non-parametric Wilcoxon estimator for continuously-distributed data and superior for inherently categorical data. The PBM MLE algorithm is well suited for use in DBM-MRMC analysis (4), because bias proportional to $1/N$ is eliminated completely by the use of jackknifing to calculate pseudovalues (25), and because the large number of leave-one-out AUC estimates needed in DBM-MRMC jackknifing requires a highly reliable curve-fitting algorithm.

As noted in the captions of Tables 2–4, the PBM algorithm's bias was always small relative to standard error and as such it is unlikely to affect statistical inference substantially. The largest ratio of bias to standard error found in any of our simulations was .44, which occurred only for the combination of large population AUC (.95), highly skewed population ROC, and small datasets (25 negative and 25 positive cases). Hence, bias was always small relative to standard error even when it was not small in an absolute sense. We also calculated root-mean-square error (RMSE) for all of our simulations but do not report it separately here because $RMSE = \sqrt{[(bias)^2 + (SE)^2]}$ and so was dominated strongly by standard error.

Although the standard-error estimates shown in Table 4 seem good on average, the corresponding Type I error rates deviated from their nominal values for that large population AUC. With the symmetric population ROC curve having $AUC = .95$, Type I error rates tended to be conservative for fewer than 100 positive and 100 negative cases. We ascribe the conservative nature of the test in this situation to our observation that the empirical distribution of AUC estimates was not close to normal, but instead was skewed by the upper bound on AUC at 1, making large differences between AUC estimates less likely. In contrast, the statistical test was only slightly conservative with fewer than 50 negative and 50 positive cases for the moderately skewed population ROC curve with $AUC = .95$, and then produced Type I error rates very close to nominal for larger numbers of cases. Although the results of these error-rate studies were too complicated overall to report in detail here, the general trend was that the statistical test ranged from conservative for symmetrical ROCs to anti-conservative for skewed ROCs. The extent to which the test was conservative or anti-conservative decreased in larger case samples, and departures from the nominal Type I error rate in various situations was found to be caused by non-normality of the test statistic, biased standard-error estimates, or both.

Additional simulations involving continuously-distributed data were performed with a variety of population PBM ROC curves that had AUCs ranging from .5 and .99 to check the stability of the algorithm when AUCs approach their limiting values of .5 and 1. Some instability in calculation of standard-error estimates by inversion of the Hessian matrix was observed for estimated AUCs very close to .5 due to the cusp-like nature of the log-likelihood function for such extreme datasets. However, the new algorithm also provides for all estimates of AUC a second standard-error estimate obtained from the *ad hoc* expression $SE\{AUC\} = 0.5\sqrt{AUC(1-AUC)(1/N_{POS} + 1/N_{NEG})}$, where N_{POS} and N_{NEG} represent numbers of positive and negative cases respectively, which we have found empirically to be accurate over a broad range of conditions (e.g., as may be demonstrated with the data in Tables 2–4) and can be compared by the algorithm's user with the inverse-Hessian estimate obtained from any single dataset to provide assurance that the latter is reasonable. No other fitting problems were observed. We note in passing that estimates of AUC were somewhat biased when population $AUC = .5$, averaging .52 in 10,000 datasets simulated for that condition; this bias was to be expected,

because the PBM does not allow ROC curves with AUCs less than .5. Therefore, the algorithm's AUC estimates and standard-error estimates must be used with caution in those rare situations where quantitative assessment or comparison of extremely low ROC curves is of practical interest. In contrast, datasets arising from the population ROC with AUC = .99 caused no problems.

When a large number of categories results from the LABROC4 categorization scheme, which occurs with very large datasets or with moderately large datasets and low population ROC curves, numerical cancellation errors in the calculation of category probabilities become potentially important, because the corresponding category boundaries then lie close to each other on the latent decision-variable axis, yielding small category probabilities. Our solution to the problem of small category probabilities has been described above. To test the validity of these solutions, a variety of simulated datasets with AUC = .6, .85 and .95, each involving at least 1000 cases, was fit with the new algorithm modified to constrain the maximum number of categories to 400 by selecting, if necessary, a well-spread subset of 399 empirical operating points from those produced by the LABROC4 categorizer. No convergence problems were encountered when this constraint was employed, and it had no measurable effect on the curve fits.

In a final simulation study, we generated a large number of 5-category datasets from the population model studied by Pan and Metz (17), which tends to produce "degenerate" (14) data. Six groups of 50,000 datasets were generated with the following combinations of actually negative (−) and actually positive (+) cases: (+/−) = (10/10), (20/20), (25/25), (50/50), (75/75) and (100/100). The percentage of degenerate datasets varied from 55% for (10/10) to 5% for (100/100). Again, no convergence problems were encountered. A more general and detailed description of the performance of the PBM with inherently categorical data is in preparation.

Discussion

The new implementation of the proper binormal ROC estimation algorithm described here provides reliable AUC estimates and curve fits with appropriate shapes in a broad variety of situations, never failing to converge for any of the millions of datasets on which it was tested. With continuous data, behavior of the algorithm's AUC estimates was approximately equivalent to that of the non-parametric Wilcoxon approach, yielding smaller RMSE though slightly higher bias. Moreover, the algorithm is robust, providing meaningful fits for inherently categorical datasets that range from small and degenerate to large and stable. Therefore, it is likely to prove superior to conventional parametric, semi-parametric and non-parametric methods, which either fail, produce unreliable fits, or yield highly biased AUC estimates in at least some such situations. In summary, the algorithm described here provides a reliable tool for ROC curve and AUC estimation in MRMC experiments that employ the DBM jackknifing technique (4), the application that motivated this research. A more complete study of the new algorithm and other approaches to ROC curve and AUC estimation for inherently categorical data is currently underway.

Acknowledgments

This work was supported by NIH grant R01 EB000863 (Kevin S. Berbaum, Principal Investigator) through a University of Chicago contract with the University of Iowa.

References

1. Metz CE, Pan X. "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *J Math Psychol* 1999;43:1–33. [PubMed: 10069933]

2. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: Part III. Initial evaluation with detection ROC data. *Acad Radiol* 2000;7:438–447. [PubMed: 10845403]
3. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Acad Radiol* 1997;4:587–600. [PubMed: 9261459]
4. 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. [PubMed: 1399456]
5. Obuchowski NA, Rockette HE. Hypothesis-testing of diagnostic-accuracy for multiple readers and multiple tests - an ANOVA approach with dependent observations. *Communications in Statistics—Simulation and Computation* 1995;24:285–308.
6. 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. [PubMed: 10803614]
7. Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: Lessons from recent experience. *Acad Radiol* 2002;9:1264–1277. [PubMed: 12449359]
8. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med* 1996;15:1807–1826. [PubMed: 8870162]
9. Gallas BD. One-shot estimate of MRMC variance: AUC. *Acad Radiol* 2006;13:353–362. [PubMed: 16488848]
10. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals — rating-method data. *J Math Psychol* 1969;6:487–496.
11. Hanley JA. Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging* 1989;29:307–335. [PubMed: 2667567]
12. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychol Bull* 1986;99:181–198. [PubMed: 3515382]
13. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med* 1998;17:1033–1053. [PubMed: 9612889]
14. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 1989;24:234–245. [PubMed: 2753640]
15. Hanley JA. The robustness of the “binormal” assumptions used in fitting ROC curves. *Med Decis Making* 1988;8:197–203. [PubMed: 3398748]
16. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: An empirical comparison of five methods. *Acad Radiol* 2004;11:980–995. [PubMed: 15350579]
17. Pan X, Metz CE. The “proper” binormal model: Parametric receiver operating characteristic curve estimation with degenerate data. *Acad Radiol* 1997;4:380–389. [PubMed: 9156236]
18. Bandos AI, Rockette HE, Gur D. Incorporating utility-weights when comparing two diagnostic systems: A preliminary assessment. *Acad Radiol* 2005;12:1293–1300. [PubMed: 16179206]
19. Obuchowski NA. Estimating and comparing diagnostic tests’ accuracy when the gold standard is not binary. *Acad Radiol* 2005;12:1198–1204. [PubMed: 16099683]
20. Edwards DC, Metz CE, Kupinski MA. Ideal observers and optimal ROC hypersurfaces in N-class classification. *IEEE Trans Med Imaging* 2004;23:891–895. [PubMed: 15250641]
21. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* 1986;21:720–733. [PubMed: 3095258]
22. Flannery, BP.; Teukolsky, SA.; Press, WH.; Vetterling, WT. *Numerical recipes in FORTRAN: The art of scientific computing*. Cambridge, England ; New York, NY, USA: Cambridge University Press; 1992.
23. Gay DM. Subroutines for unconstrained minimization using a model trust-region approach. *ACM Transactions on Mathematical Software* 1983;9:503–524.

24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845. [PubMed: 3203132]
25. Efron, B.; Tibshirani, R. An introduction to the bootstrap. New York, NY, USA: Chapman & Hall; 1993.
26. Rao, CR. Linear statistical inference and its applications. New York, NY, USA: Wiley; 1965.
27. Papoulis, A. Probability, random variables, and stochastic processes. New York, NY, USA: McGraw-Hill; 1984.
28. Campbell, SL.; Meyer, CD. Generalized inverses of linear transformations. New York, NY, USA: Dover Publications; 1991.
29. Rao, CR.; Mitra, SK. Generalized inverse of matrices and its applications. New York, NY, USA: Wiley; 1971.
30. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989;9:190–195. [PubMed: 2668680]
31. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996;201:745–750. [PubMed: 8939225]
32. Metz C, Kronman H. Statistical significance tests for binormal ROC curves. *J Math Psychol* 1980;22:218–243.
33. Kupinski MA, Edwards DC, Giger ML, Metz CE. Ideal observer approximation using Bayesian classification neural networks. *IEEE Trans Med Imaging* 2001;20:886–899. [PubMed: 11585206]
34. Glantz, SA. Primer of biostatistics. New York, NY, USA: McGraw-Hill, Medical Pub. Div; 2002.
35. Metz CE. Basic principles of ROC analysis. *Seminars Nucl Med* 1978;8:283–298.
36. Shiraishi J, Katsuragawa S, Ikezoe J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am J Roentgenol* 2000;174:71–74. [PubMed: 10628457]
37. Bamber D. Area above ordinal dominance graph and area below receiver operating characteristic graph. *J Math Psychol* 1975;12:387–415.

Figure 1.

Proper binormal ROC fits (solid curves), conventional binormal ROC fits (long-dashed curves) and Wilcoxon trapezoidal interpolations (short dashes) for four of the problematic datasets reported in (2). Empirical operating points calculated directly from the data are represented by black dots. Although binomial ± 2 SE error bars are omitted here for clarity, the curves fit by the new algorithm lie well within them at each empirical operating point.

Figure 2.

Fits for a large dataset reported by Kupinski et al. (33). Empirical operating points are represented by black diamonds. Vertical and horizontal binomial ± 2 SE error bars for each empirical operating point were calculated by the method described in the text.

Figure 3.

Fits for two large datasets reported by Shiraishi et al. (36). Empirical operating points are represented by black diamonds. Vertical and horizontal binomial $\pm 2SE$ error bars for each empirical operating point were calculated by the method described in the text. Each panel is labeled with the reader number used in the original paper. These datasets were selected for display and discussion here because their fits seem to depart from the empirical points.

Figure 4. Population PBM ROC curves used to generate the continuously-distributed simulated datasets reported in this paper. The AUC values are .60000 for the lowest triad of curves, .85000 for the middle curves, and .95000 for the highest. The PBM parameter values associated with these population curves are shown in Table 1.

Parameter values used in the simulations
AUC and parameter values of the population ROC curves used in the simulation studies. The corresponding plots are shown in Fig. 4.
These AUC values and curve shapes bracket the ranges encountered in most medical-imaging studies. The parameter values are specified to six significant digits because this precision is required for accurate calculation of some AUC values.

AUC value	Curve shape (skew)					
	High		Medium		Symmetrical	
	da	c	da	c	da	c
.60000	.100000	-.154357	.358000	-.046907	.358287	0.0
.85000	1.24365	-.428571	1.46571	-.140000	1.46574	0.0
.95000	.0999998	-.853738	2.32400	-.292369	2.32618	0.0

Table 2

Results of the simulations for AUC = .60000

The value shown at the top of each column represents the number of actually positive or negative cases (i.e., half of the total number of cases) used in each simulated experiment. The rows within each curve-shape block indicate: $\overline{AUC}$, average of the estimated AUC value; SE , empirical standard deviation of the estimated AUC values; $\widehat{SE}$, average of the estimated standard errors determined from the individual AUC estimates; SE/SE_w , ratio of the empirical standard deviation of the MLE AUC estimates to that of the Wilcoxon AUC estimates. Each table entry was obtained from 20,000 datasets to reduce uncertainties to less than three digits, so no confidence intervals are shown. The largest bias/standard-deviation ratios (.27 for the symmetric curve, .40 for the highly skewed curve) were observed for the datasets comprised of 25 actually negative and 25 actually positive cases. The estimated ratio of MLE to Wilcoxon standard error carries an uncertainty of approximately $\pm 1\%$; this precision is related to the number of datasets, rather than to the size of each dataset, so it is approximately the same for all numbers of cases.

ROC curve shape (skew)		Number of actually positive or negative cases					
		25	50	75	100	200	300
Symmetric (no skew)	$\overline{AUC}$	.618	.608	.604	.603	.6003	.600
	SE	.070	.051	.042	.038	.027	.023
	$\widehat{SE}$	.071	.052	.043	.038	.027	.022
	SE/SE_w	.73	.81	.85	.89	.94	.96
sMedium skew	$\overline{AUC}$	.621	.610	.606	.604	.602	.600
	SE	.069	.050	.041	.036	.026	.022
	$\widehat{SE}$	.070	.051	.042	.037	.026	.022
	SE/SE_w	.85	.89	.90	.92	.94	.95
High skew	$\overline{AUC}$	.627	.617	.613	.612	.608	.606
	SE	.068	.049	.040	.034	.024	.020
	$\widehat{SE}$	.069	.050	.041	.035	.025	.020
	SE/SE_w	.84	.87	.86	.87	.86	.86

Table 3

Results of the simulations for AUC = .85000

The notation used here is defined in the caption of Table 2. For the smallest datasets (25 actually negative and 25 actually positive cases), the ratio of bias to standard deviation is .01 for the worst symmetrical population ROC and .36 for the worst highly-skewed ROC. Again, the ratio of MLE to Wilcoxon standard error has an uncertainty of approximately $\pm 1\%$.

ROC curve shape (skew)		Number of actually positive or negative cases							
		25	50	75	100	200	300	400	500
Symmetric (no skew)	AUC	.855	.852	.851	.851	.851	.8505	.8503	.8503
	SE	.053	.038	.031	.026	.018	.015	.013	.012
	$\hat{SE}$	.052	.037	.030	.026	.018	.015	.013	.012
	SE/SE _W	.98	1.00	.99	1.00	.99	.99	1.00	1.00
Medium skew	AUC	.855	.852	.851	.8506	.8504	.8502	.8502	.8502
	SE	.052	.038	.031	.027	.019	.016	.013	.012
	$\hat{SE}$	.051	.037	.031	.027	.019	.015	.013	.012
	SE/SE _W	.95	.99	.99	1.00	1.00	1.00	.99	.98
High skew	AUC	.867	.860	.857	.856	.853	.852	.8516	
	SE	.047	.034	.028	.024	.017	.014	.012	
	$\hat{SE}$	.046	.033	.027	.024	.017	.014	.012	
	SE/SE _W	.84	.86	.86	.86	.88	.89	.88	

Results of the simulations for AUC = .95000
The notation used here is defined in the caption of Table 2. For the smallest datasets, the ratio of bias to standard deviation is .035 for the worst symmetrical population ROC and .44 for the worst highly-skewed ROC. As in Tables 2 and 3, the estimated ratio of MLE to Wilcoxon standard error has an uncertainty of approximately $\pm 1\%$.

Table 4

ROC curve shape (skew)		Number of actually positive or negative cases					
		25	50	75	100	200	300
Symmetric (no skew)	AUC	.951	.951	.9505	.9504	.9503	.9501
	SE	.028	.019	.016	.014	.010	.008
	$\hat{SE}$	.029	.020	.016	.014	.010	.008
	SE/SE _W	.96	.95	.95	.95	.95	.95
Medium skew	AUC	.954	.951	.9503	.9501	.9500	.9500
	SE	.028	.021	.018	.015	.011	.009
	$\hat{SE}$	.028	.021	.017	.015	.011	.009
	SE/SE _W	.96	.95	.95	.95	.95	.95
High skew	AUC	.961	.957	.956	.955	.953	.952
	SE	.025	.018	.015	.013	.009	.008
	$\hat{SE}$	.024	.018	.015	.013	.009	.008
	SE/SE _W	.96	.96	.96	.96	.95	.95