Published in final edited form as:

Acad Radiol. 2010 August ; 17(8): 969–979. doi:10.1016/j.acra.2010.03.020.

The Effect of Two Priors on Bayesian Estimation of “Proper” Binormal ROC Curves from Common and Degenerate Datasets

Richard M. Zur, B.Sc., Lorenzo L. Pesce, Ph.D., and Yulei Jiang, Ph.D.

Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, MC2026, Chicago, IL 60637

Abstract

Rationale and Objectives—We showed previously that maximum-likelihood (ML) and Bayesian (with a flat prior on a common parameterization of the model) estimates of “proper” binormal receiver operating characteristic (ROC) curves produce similar results. We propose a new prior that is flat over the area under the ROC curve (AUC) and investigate its effect on the Bayesian estimates.

Materials and Methods—In two simulation studies we compared Bayesian estimation of the AUC with the two prior probability distributions against ML estimation in terms of root-mean squared error (RMSE) and the coverage of 95% confidence (or credible) intervals (both abbreviated CIs). In the first study we simulated categorical data that tend to be “well-behaved” and produce ROC curve estimates that most would consider reasonable. In the second study we simulated coarsely discretized categorical data that often included so-called degenerate datasets that cause the ML estimate to be the perfect ROC curve.

Results—For the well-behaved datasets, all three AUC estimates were similar in terms of RMSE and 95% CI coverage. For the coarsely discretized datasets, the RMSE of ML was consistently greater than that of Bayesian estimation and the 95% CI coverage of ML estimation was consistently below nominal whereas the 95% CI coverage of Bayesian estimation was consistently equal to, or greater than, nominal.

Conclusion—Bayesian estimation with a flat prior on the AUC can provide reasonable inference from datasets with coarsely categorized data that are prone to be degenerate, and produce results similar to other estimation methods on well-behaved datasets.

Keywords

ROC curves; Bayesian analysis; degenerate data

© 2010 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.

Address correspondence to: Richard M. Zur, Department of Radiology, MC2026, The University of Chicago, 5841, South Maryland Avenue, Chicago, IL, 60637, Telephone: (773) 834-5094, FAX: (773) 702-0371, zur@uchicago.edu.

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.

1. Introduction

Receiver operating characteristic (ROC) analysis is a fundamental method for the evaluation of diagnostic accuracy (1-3). An ROC curve is a plot of true positive fraction (TPF, or sensitivity) versus false positive fraction (FPF, or 1-specificity). The conventional binormal model for ROC analysis provides satisfactory ROC curve fits in a wide variety of practical situations (4-7). However, except for ROC curves that are symmetrical with respect to the negative 45° line in the ROC plot, the conventional binormal model produces ROC curve estimates that contain “hooks” (i.e., a change in the ROC curve curvature, e.g., from convex to concave, which for the conventional binormal model implies that a portion of the ROC curve falls below the “guessing line” defined by $TPF = FPF$). These ROC curve estimates are considered unsatisfactory because hooks indicate diagnostic accuracies that are worse than guessing (8-10). ROC models that describe ROC curves guaranteed to have a monotonically decreasing slope are known as “proper” ROC models (3). In this paper, we focus on the so-called “proper” binormal model (11, 12).

Previously (13), we compared ML and Bayesian estimates of “proper” binormal ROC curves. The Bayesian estimates were based on a prior probability distribution that is flat (i.e., constant) over the most common parameterization of the proper binormal model (11). Prior probability distributions are a well-known characteristic of Bayesian estimation and they incorporate information obtained independently from the data at hand (14, 15). Because of that, priors are expected either to improve estimations or to bias them. We showed (13) that the Bayesian and ML estimates are similar in terms of root mean squared error (RMSE) for the area under the ROC curve (AUC), TPF values at fixed FPF values and FPF values at fixed TPF values. In this paper, we evaluate the effect on Bayesian estimation of proper binormal ROC curves of a new prior that is not flat over the most common parameterization of the proper binormal model but, rather, flat over the AUC values. We refer to both of these flat priors as low-information because they do contain information (as will be demonstrated later through re-parameterization) and can affect ROC analysis. Our motivation was that whereas a prior flat on the curve parameters is expected to influence minimally the estimation of the curve parameters, it is probably more desirable to influence minimally the estimation of the AUC—the most commonly reported ROC summary index in radiological research (16). It is not possible to impose a prior that is simultaneously flat on both the curve parameters and the AUC because the curve parameters and the AUC are non-linear functions of each other (17). Therefore, a trade-off is unavoidable. Although the effect of prior information or preconceived views is usually not discussed in the context of ROC curve estimation, a review of the radiological literature that reports ROC analysis shows that prior experience and belief do seem to influence our understanding and acceptance of ROC curve estimates (16).

2. Background

2.1 The “proper” binormal ROC model

The proper binormal ROC model introduced by Metz et al. (11, 12) is usually specified by two parameters: d_a and c , which define the signal-present and signal-absent distributions of cases with respect to a latent decision variable, v (for details refer to Metz *et al.* (11)). The

minimum of the AUC value (0.5) occurs when $d_a = 0$ and $c = 0$, and the AUC value increases when either d_a increases toward infinity or $|c|$ increases towards 1, or both. The ROC curve is symmetrical around the negative 45° line in the ROC plot when $c = 0$, whereas the ROC curve is skewed toward the left (i.e., toward the ordinate) when $c < 0$ and toward the right with respect to the negative 45° line in the ROC plot (i.e., toward the line defined by $\text{TPF} = 1$) when $c > 0$.

In this paper we treat all ROC data as ordinal-categorical, because the limited precision of real-world quasi-continuous data implies that they can be mapped to some ordinal-categorical data without loss of information relevant to ROC analysis. We partition the latent decision variable axis into I categories with $I-1$ cut-points, v_{ci} , $i = 1, 2, \dots, I-1$, hereafter denoted by $\{v_{ci}\}$ (18-20). If necessary, we reduce the number of categories to $I = 20$ using the LABROC5 algorithm (21), which appears to affect the bias and variance of ML estimates negligibly as long as at least 20 categories are used (12, 21). As is customary, we consider these cut-points as nuisance parameters because we are more interested in estimating the ROC curve than in estimating the cut-points. Therefore, the likelihood function—the probability of the data conditional on the model—assumes that the ROC data are obtained through multinomial sampling (21).

2.2 Maximum-likelihood estimation of the “proper” binormal ROC model

Let us denote the likelihood function by $L(d_a, c, \{v_{ci}\}; \mathbf{D}) = p(\mathbf{D} | d_a, c, \{v_{ci}\})$, where the semi-colon indicates that the likelihood function is dependent, but not conditional, on the data, $\mathbf{D}$. The ML estimate is the set of parameter values, d_a , c and $\{v_{ci}\}$ that maximizes $L(\bullet)$. Uncertainty in d_a , c and $\{v_{ci}\}$ can be estimated from the inverse of the Fisher information matrix (22). Uncertainty in other quantities of interest, e.g., the AUC, TPF and FPF, can be estimated from the Fisher information matrix using the delta method (17). An extensive discussion on this likelihood function, the ML estimate and the estimate of the variance-covariance matrix can be found elsewhere (12).

2.3 Bayesian estimation of the “proper” binormal ROC model

From Bayes' rule, the posterior probability distribution of the proper binormal model parameters is given by

$$p(d_a, c, \{v_{ci}\} | \mathbf{D}) = \frac{\mathbf{L}(\mathbf{d}_a, \mathbf{c}, \{\mathbf{v}_{ci}\}; \mathbf{D}) \mathbf{p}(\mathbf{d}_a, \mathbf{c}, \{\mathbf{v}_{ci}\})}{\mathbf{p}(\mathbf{D})}, \quad (1)$$

where $p(d_a, c, \{v_{ci}\})$ is the prior probability distribution and $p(\mathbf{D})$ is the marginal likelihood (14). The marginal likelihood can be considered as a normalization constant, which does not affect most estimates of the posterior probability distribution (14). Therefore, Bayesian estimation focuses on the probability of the model given the data, i.e., the posterior probability distribution, rather than on the probability of the data given a model, i.e., the likelihood function. However, as shown in equation (1), a prior probability distribution is required to estimate the posterior probability distribution. It is not always clear what the best or even a reasonable prior probability distribution is. Furthermore, we are often interested in indices that are different from, or in addition to, the parameters that we estimate directly. In

such instances, even after necessary transformations (17), we sometimes find that a prior probability distribution that is reasonable for one set of parameters does not appear to be reasonable for other parameters of interest. Moreover, to estimate a high dimensional, non-standard, probability distribution, as is often required with Bayesian estimation, can be computationally challenging. Here, we use Markov chain Monte Carlo (MCMC) algorithms for that purpose (23).

2.4 Markov chain Monte Carlo estimation of the Bayesian posterior probability distribution

MCMC is a common approach to obtaining random samples from probability distributions, even of large dimensions (e.g., > 10) (23-25). MCMC algorithms are well-suited for Bayesian estimation of ROC curves because the estimation can involve a large number of dimensions due to potentially large numbers of the cut-points, $\{v_{ci}\}$. With MCMC algorithms, integrals over probability distributions necessary for calculating summary quantities such as the mean, median and credible intervals (similar to confidence intervals in ML estimation, both abbreviated CI) can be approximated with summations over random samples (which the MCMC samples approximate) from the target probability distribution. The MCMC method also naturally enforces bounds, such as the bounds in the values of the AUC, TPF and FPF, because the algorithm can produce numerical samples only within the correct bounds. We used a slice sampling MCMC algorithm (26) as we did in a previous study (13).

We express our Bayesian estimation tasks in terms of the parameters d_a , c and $\{v_{ci}\}$ and transform the results to obtain estimates of other quantities of interest. One advantage of this approach is that we can easily compare the effect of priors of different functional forms without having to modify the machinery and procedures of our Bayesian estimation except for re-defining the prior probability distribution.

2.5 Degenerate datasets compatible with the perfect ROC curve

When all empirical operating points (i.e., {FPF, TPF} pairs) of an ROC dataset either have only values of FPF = 0 or have only values of TPF = 1, or both, we call the dataset degenerate and compatible with the perfect ROC curve (1). These datasets are “degenerate” because ML estimation of most ROC models that are based on the categorical data-modeling approach (19, 20), including both the conventional and the proper binormal models, fit these datasets with an ROC curve composed of straight line segments (10). These degenerate datasets are “compatible with the perfect ROC curve” because the perfect ROC curve is an exact fit to them. Throughout this paper, we will use the term degenerate dataset to refer to only those degenerate datasets that are compatible with the perfect ROC curve.

ML estimation of the proper binormal ROC model on a degenerate dataset produces a result defined by either $|c| = 1.0$ and any arbitrary value of d_a , or $d_a = \infty$ and any arbitrary value of c , both with AUC = 1.0 (typically an overestimate of the population AUC value) and zero variance. On the other hand, the empirical ROC curve connects the operating points with straight lines, giving rise to an ROC curve with a smaller AUC value (usually an underestimate of the population AUC value) (27).

3. Materials and Methods

3.1 The prior probability distributions

We define the prior probability distributions as functions of d_a , c and $\{v_{ci}\}$. We further construct the prior probability distributions independently with respect to d_a and c , and with respect to $\{v_{ci}\}$, i.e., $p(d_a, c, \{v_{ci}\}) = p(d_a, c) \times p(\{v_{ci}\})$, which does not imply independence in the posterior probability distribution because the likelihood function involves explicit dependencies among all parameters. We set $p(\{v_{ci}\}) = k$, where k is a constant, because in general we do not know what probability is associated with different $\{v_{ci}\}$ values. This construct of the prior did not appear to affect our results in a previous study (13). We call $p(\bullet) = k$ a “flat” prior.

A flat prior on the curve parameters d_a and c , i.e., $p(d_a, c) = k$, implies that the posterior probability distribution is proportional to the likelihood function (both expressed in terms of d_a and c) and, therefore, that the Bayesian maximum *a posteriori* (MAP) estimate is equal to the ML estimate of these parameters. Thus, with a flat prior on d_a and c , any differences between the ML and MAP estimates are the result of differences between the methods of ML and Bayesian estimations. However, as Fig. 1a) illustrates, there are greater numbers of d_a and c value combinations that produce large AUC values than there are that produce small AUC values. Therefore, the probability density distribution of the AUC value is not flat given the flat prior on d_a and c (Fig. 1b).

Therefore, to explore the possibility of an alternative prior more consistent with our use of ROC analysis, and to evaluate the effect of priors on Bayesian estimation of proper binormal ROC curves, we propose a new prior probability distribution that is marginally flat on the AUC and flat on c (conditionally on AUC), which we refer to as the “flat marginalized AUC” or simply the “flat AUC” prior. We define this prior to be flat over c because c is related to the shape of the ROC curve. We express this prior also in terms of d_a and c , and give the mathematical details that define this prior in the Appendix. The log of the flat AUC prior is shown in Fig. 2. It has a maximum at $d_a = 0$ and $c = 0$ and decreases asymptotically as d_a approaches zero (when $c = 0$) or infinity, and as $|c|$ approaches 1, or both. Therefore, this prior probability distribution reduces the probability of d_a and c value combinations that correspond to large AUC values relatively to the ones that produce small AUC values.

3.2 Simulation studies

We compared the ML and Bayesian estimates of proper binormal ROC curves with two kinds of simulated ROC datasets: one that produced categorical datasets defined over a relatively large number of categories and one that produced coarsely categorized datasets that are prone to be degenerate (10). The first kind of datasets is more common in practice and we refer to them as “well-behaved” datasets in this paper. We did not analyze degenerate datasets with Bayesian estimation using a flat prior on d_a and c because in such cases the posterior probability does not go to zero as d_a increases towards infinity or as $|c|$ increases towards 1.0, therefore its integral is not finite, causing MCMC sampling to not converge. We compared AUC estimates obtained with the ML and Bayesian approaches. We simulated well-behaved datasets from nine “population” ROC curves that broadly cover

the range of ROC curves reported in the literature, with AUC values of 0.60, 0.85 or 0.95 and with the curve being of no skewness, “slight” skewness or “moderate” skewness, as described in detail elsewhere (12). We simulated coarsely categorized datasets from three population ROC curves with AUC values of 0.60, 0.85 or 0.95 and with the curve being of slight skewness.

From each population ROC curve we simulated 1,000 ordinal-categorical datasets, each composed of 25 signal-present and 25 signal-absent cases. We studied ROC datasets with a relatively small number of cases because the differences between estimation methods are greater with small datasets and because degenerate datasets are more likely to occur in small datasets. To obtain coarsely categorized datasets we simulated three-category datasets using two fixed v_{ci} values, which gave rise to at most two operating points. For the population ROC curves of AUC = 0.60, we used $v_{c1} = -1.5$ and $v_{c2} = 1.5$, $v_{c1} = -1.7$ and $v_{c2} = 1.7$, and $v_{c1} = -1.8$ and $v_{c2} = 2.0$ to yield approximately 15%, 50% and 70%, respectively, of the total datasets being degenerate. These v_{ci} values were, for the population ROC curves of AUC = 0.85: $v_{c1} = -1.0$ and $v_{c2} = 1.0$, $v_{c1} = -1.5$ and $v_{c2} = 1.5$, and $v_{c1} = -2.0$ and $v_{c2} = 2.0$, respectively; and for the population ROC curves of AUC = 0.95: $v_{c1} = -0.7$ and $v_{c2} = 0.4$, $v_{c1} = -1.2$ and $v_{c2} = 0.65$, and $v_{c1} = -1.5$ and $v_{c2} = 0.85$.

Our simulation yielded 297 (out of a total of 9,000) simulated coarsely categorized datasets that had a single operating point (i.e., all cases fell into two categories) with a value of TPF = 1 or FPF = 0 (i.e., not consistent with a perfect ROC curve.) For these cases we used an *ad hoc* algorithm to obtain the ML estimate of the AUC and its standard error (12) (note that an infinite number of ROC curves with equal likelihood pass through this single operating point).

We obtained ML estimates with the PROPROC algorithm (12) and Bayesian estimates with a custom-written computer program (13). To obtain Bayesian estimates of the AUC value, we transformed the MCMC samples of the posterior probability distribution of d_a and c into a set of corresponding samples of the posterior probability distribution of the AUC values. We obtained two summary point estimates from the Bayesian posterior probability estimates: the mean of the posterior probability distribution calculated as the arithmetic average of the estimated posterior probability distribution, and the MAP calculated as the maximum of the estimated posterior probability distribution after smoothing it using a Gaussian kernel with an automatically selected bandwidth (28). We estimated the Bayesian 95% CIs by explicitly excluding 2.5% of the MCMC samples from each tail of the estimated posterior probability distribution.

3.3 Analysis of simulation study datasets

We compared the bias, standard deviation and root mean-squared error (RMSE) of the ML and Bayesian AUC estimates. The RMSE—the square root of the sum of bias squared and standard deviation squared—is a measure of the overall error of an estimator. We also compared the coverage of the 95% CIs of the AUC estimate, i.e., the proportion of all estimated intervals that enclosed the population AUC value, against the nominal value of 95%.

4. Results

Figure 3 shows example ML and Bayesian ROC curve estimates from a well-behaved dataset simulated from the population ROC curve with $AUC = 0.60$ and no skewness ($d_a = 0.36$ and $c = 0.0$, panels a and b) and from a degenerate dataset simulated from the population ROC curve with $AUC = 0.60$ and slight skewness ($d_a = 0.358000$ and $c = -0.046907$, the six significant digits was to ensure $AUC = 0.60$ to six significant digits, $v_{c1} = -1.5$ and $v_{c2} = 1.5$, panels c and d). The Bayesian estimates are shown as 95%, 90%, 75% and 50% iso-probability contours. The ML estimates of the AUC value were 0.70 for the well-behaved dataset and 1.00 for the degenerate dataset. The Bayesian mean-posterior estimates of the AUC value were 0.73 (flat prior on d_a and c) and 0.71 (flat prior on the AUC) for the well-behaved dataset, and 0.79 (flat prior on the AUC) for the degenerate dataset. The Bayesian estimates of the AUC value were similar to the ML estimate on the well-behaved dataset, but clearly closer to the population AUC value than the ML estimate on the degenerate dataset.

Figure 4 summarizes ML and Bayesian estimates of the AUC value for the well-behaved datasets. The bias in the ML estimates ranges from 0.00 to 0.03, which is consistent with previous results (12). The bias in the Bayesian estimates with the flat prior on d_a and c ranges from -0.02 to 0.05, whereas the bias of the Bayesian estimates with the flat prior on the AUC ranges from -0.03 to 0.03. The differences between the RMSE of all estimates are small, except that the RMSEs of the Bayesian mean-posterior estimates with the flat prior on the AUC are smaller for the three population ROC curves with $AUC = 0.60$ but slightly larger for five of the other six population ROC curves compared with the RMSEs of the other two estimation methods, and are more uniform across all population ROC curves compared with the other two estimation methods.

Comparisons of the coverage and width of the 95% CIs in AUC estimates between the ML and Bayesian estimation are shown in Fig. 5. The width of the 95% CIs is important because it is desirable for a 95% CI estimate that produces at least nominal coverage to be as narrow as possible (29). The average widths between all three methods are similar. In 63 of a total of 3,000 datasets from the population ROC curves of $AUC = 0.95$, the logit transformation that we applied to ensure correctly bounded ML estimate of the 95% CIs resulted in wide CI estimates that are practically uninformative (13). The coverage of the ML estimates and Bayesian estimates with the flat prior on d_a and c is similar. The coverage of the Bayesian estimates with the flat prior on the AUC appears to be less than nominal for the population ROC curves of $AUC = 0.85$ and $AUC = 0.95$ and greater than nominal for the population ROC curves of $AUC = 0.60$. However, all differences in the coverage of the 95% CIs appear to be small.

Figure 6 compares ML and Bayesian estimates of the AUC from coarsely categorized datasets that are prone to be degenerate. The bias of the ML estimates ranges from 0.00 to 0.09 when the probability of degenerate datasets is 15% and increases to 0.03 to 0.31 when the probability of degenerate datasets is 70%. In comparison, the bias of the Bayesian MAP estimates with the flat prior on the AUC ranges from -0.01 to 0.02 when the probability of degenerate datasets is 15% and is -0.06 to 0.02 when the probability of degenerate datasets

is 70%, whereas the bias of the Bayesian mean-posterior estimates with the flat prior on the AUC ranges from -0.03 to 0.05 when the probability of degenerate datasets is 15% and is -0.07 to 0.09 when the probability of degenerate datasets is 70%. The RMSE of the ML estimates ranges from 0.03 to 0.19 when the probability of degenerate datasets is 15% and ranges from 0.04 to 0.35 when the probability of degenerate datasets is 70%. On the other hand, the RMSE of the Bayesian MAP estimates with the flat prior on the AUC ranges from 0.04 to 0.09 when the probability of degenerate datasets is 15% and ranges from 0.03 to 0.10 when the probability of degenerate datasets is 70%, whereas the Bayesian mean-posterior estimates with the flat prior on the AUC ranges from 0.04 to 0.08 when the probability of degenerate datasets is 15% and ranges from 0.04 to 0.10 when the probability of degenerate datasets is 70%. Therefore, the RMSE of the Bayesian estimates of the AUC value is considerably smaller than that of the ML estimates when degenerate datasets are present.

Figure 7 compares the coverage and average width of the 95% CIs for the AUC estimate between ML and Bayesian estimation from coarsely categorized datasets that are prone to be degenerate. The coverage of the ML estimates ranges from 0.85 to 0.86 when the probability of degenerate datasets is 15%, and reduces to a range of 0.16 to 0.41 when the probability of degenerate datasets is 70%. In comparison, the coverage of the Bayesian estimates with the flat prior on the AUC ranges from 0.95 to 0.99 when the probability of degenerate datasets is 15%, and remains at a range of 0.97 to 1.00 when the probability of degenerate datasets is 70%. When the probability of degenerate datasets is 15%, the average widths of the Bayesian 95% CIs are similar to the average widths of ML 95% CIs. As the probability of degenerate datasets increases the average widths of ML 95% CIs decrease because the proportion of ML 95% CIs with zero width (the perfect ROC curve fit) increases, but the average widths of Bayesian 95% CIs increase because the widths of Bayesian 95% CIs for degenerate datasets are generally larger than those for well-defined datasets. Figure 8 shows the same 95% CI results plotted such that the AUC estimates are sorted in increasing order. This figure shows that the Bayesian 95% CI estimates tend to be informative even though they are often wide estimates, whereas the ML estimates are often too wide to be informative. This figure also shows that the Bayesian estimates apparently form two bands that correspond to degenerate (wider CIs) and well-behaved (narrower CIs) datasets.

5. Discussion

This study underscores the differences between ML and Bayesian estimations. ML estimation yields the ROC curve most likely to have produced the observed data and, for the degenerate datasets that we analyzed here, that ROC curve is the perfect ROC curve. On the other hand, Bayesian estimation produces the most probable estimates given both the observed data and the prior information or expectation. The presence of prior information in Bayesian estimation is a fundamental difference from ML estimation and we have shown that it can lead to more reasonable ROC curve estimates for degenerate datasets.

To our knowledge, this is the first method that can fit sensible ROC curves from degenerate datasets. In principle, any prior probability distribution that approaches zero faster than the growth of the likelihood function as AUC increases towards 1.0 (i.e., d_a increases towards infinity or $|c|$ increases towards 1.0, or both) would produce ROC curve estimates with AUC

< 1.0 for degenerate datasets. We have chosen a prior with all AUC values equally probable because it seems to be a reasonable assumption for a prior distribution.

The Bayesian implementation of the proper binormal model was able to fit degenerate datasets with ROC curves of $\text{AUC} < 1.0$ because of the information contained in the flat AUC prior. In fact, both prior probability distributions described in this paper contain information, despite being flat in one parameterization. However, upon transformation, we observe that one is strongly peaked at $\text{AUC} = 1.0$ and the other is strongly peaked at $d_a = c = 0$. In this sense we expect them to be extreme examples of low-information priors, and that other low-information priors to provide estimates somewhere in between these two. However, the effect of more complex and potentially more informative prior probability distributions on the bias, standard deviation, RMSE and coverage of the 95% CIs of Bayesian estimates cannot be easily predicted and must be evaluated in simulation studies similar to what we have performed here in evaluating the flat prior on the AUC.

A major difference between the two priors we investigated is that the prior that is flat on d_a and c is not a proper prior (i.e., it is not a true probability because it has a divergent integral), whereas the flat AUC prior is proper. However, this is not the only reason for better performance of the flat AUC prior on degenerate datasets. For example, the flat d_a and c prior could be made proper by truncating it at some d_a value. The results of Bayesian estimation with a truncated, flat d_a and c prior would be similar to those of ML estimation, with the MAP of $\text{AUC} = 1$ but the mean-posterior estimate of $\text{AUC} < 1$ and dependent on the specific location of the truncation — the inherently arbitrary nature of these choices is part of the reason why we did not pursue this type of prior.

For experiments where degenerate data are likely to be found (or have been found) — e.g., in a multi-reader, multi-case observer study where each reader reads a small number of cases and uses a categorical scoring system — our simulation results suggest that performing a Bayesian analysis with the proper binormal model and a marginally flat prior on the AUC is likely superior to ML estimation. We recommend (this is the scenario that we simulated) that users use a consistent approach for all ROC curve fittings, whether the data are degenerate or not, so as not to introduce bias into the data analysis. On the other hand, although our method appears to provide reasonable estimates of ROC curves for both non-degenerate and degenerate data, experimenters should design studies that avoid producing degenerate data altogether (10).

Obuchowski proposed to use Monte Carlo simulations to find a reasonable lower bound of the 95% CI when the non-parametric (i.e., trapezoidal) AUC estimate is 1.0 (30). Strictly speaking, their method applies only to datasets that have a single operating point of $\text{FPF} = 0$ and $\text{TPF} = 1$. However, their method could be generalized to other types of degenerate datasets. Although their method is based on the non-parametric AUC estimate, it requires several parametric modeling assumptions to which the estimate of the lower bound of 95% CI can be expected to be sensitive (30). Moreover, it is well-known that the non-parametric AUC estimate underestimates the area with categorical data, especially when the categories are small in number and/or poorly spaced along the ROC curve (3, 31), which could affect the estimate of the lower bound of the 95% CI using their method.

We considered ROC datasets that consisted of only 25 signal-present and 25 signal-absent cases. As datasets increase in size, the relative importance of the prior distribution is reduced with respect to the likelihood, therefore we expect even smaller differences between ML and Bayesian estimation than what we report here. ROC datasets smaller than what we considered here carry large uncertainties and potential biases; their analysis is best done carefully with respect to a specific data source and not in general terms.

In our study we found that it is difficult to assign a prior to the cut-points, $\{v_{ci}\}$. In this study we used the prior probability distribution of $p(\{v_c\}) = k$ because we found in a previous study that this prior did not adversely and measurably affect the bias, standard deviation, RMSE and coverage of the 95% CIs of Bayesian estimates (13). However, if a more informative prior probability distribution for $\{v_{ci}\}$ is identified, it is possible that our estimates will improve.

Because of space limitation, we did not present results in the comparison of estimated ROC curves (specifically, TPF values conditional on fixed FPF values and FPF values conditional on fixed TPF values). Those results are similar to the results of the AUC estimates: for well-behaved datasets, the RMSEs of the three methods are similar, whereas for degenerate datasets, the RMSEs of Bayesian estimates with the flat prior on the AUC are smaller than those of the ML estimates.

In conclusion, for well-behaved datasets Bayesian estimation of proper binormal ROC curves with the flat prior on d_a and c and the flat prior on the AUC both produce similar results as ML estimation in terms of RMSE, but the coverage of the 95% CIs of Bayesian estimation with the flat prior on the AUC sometimes appears to be slightly less than nominal. On the other hand, for data sources that produce coarsely categorized data that are prone to be degenerate, similar to those studied in this paper, Bayesian estimation with the flat prior on the AUC will likely produce estimates with better statistical properties than ML estimation. Moreover, the Bayesian method is one of few methods that can produce an ROC curve estimate that is not the perfect ROC curve when applied to a degenerate dataset. These features are important because degenerate datasets can occur in observer performance studies that employ a limited number of action-related categories.

Acknowledgments

We thank Drs. Charles Metz and Vanja Dukic for many thoughtful discussions and comments on this manuscript.

This work was supported in part by the National Cancer Institute of the National Institutes of Health (NIH) through grants R01 CA092361 and P30 CA14599, by the National Center for Research Resources of NIH through grant S10 RR021039, and by the National Institute of Biomedical Imaging and Bioengineering of NIH through grant R01 EB000863 (Kevin S. Berbaum, PI).

Appendix

We derive a joint prior probability distribution expressed in terms of d_a and c that corresponds to a prior probability distribution that is marginally flat over the AUC. We start with a uniform, properly normalized ($0.5 < \text{AUC} < 1.0$), prior probability distribution over the AUC,

$$p(AUC)=2. \quad (1)$$

For a given AUC value, it is possible to show that c is bound between $-\tan(2\pi(AUC - 0.5))$ and $+\tan(2\pi(AUC - 0.5))$. Therefore, the normalized probability of c conditional on the AUC is

$$p(c|AUC)=\frac{1}{2\tan(2\pi(AUC - 0.5))}, \quad (2)$$

and the joint probability is

$$p(AUC, c)=p(c|AUC)p(AUC)=\frac{1}{\tan(2\pi(AUC - 0.5))}. \quad (3)$$

To transform this prior probability distribution expressed in terms of the AUC and c to a prior probability distribution expressed in terms of d_a and c , we use the determinant of the Jacobian:

$$J(d_a, c)=\left| \begin{array}{cc} \frac{\partial AUC}{\partial d_a} & \frac{\partial AUC}{\partial c} \\ \frac{\partial c}{\partial d_a} & \frac{\partial c}{\partial c} \end{array} \right| = \frac{\partial AUC}{\partial d_a} \frac{\partial c}{\partial c} - \frac{\partial AUC}{\partial c} \frac{\partial c}{\partial d_a} = \frac{\partial AUC}{\partial d_a} = \frac{1}{2\sqrt{\pi}} e^{-\frac{d_a^2}{4}} \left[1 - 2\Theta \left(\frac{\frac{c^2-1}{c^2+1}d_a}{\sqrt{2-2\left(\frac{c^2-1}{c^2+1}\right)^2}} \right) \right], \quad (4)$$

where Θ is the cumulative standard normal function. Therefore, the prior probability distribution over d_a and c that corresponds to a prior probability distribution that is marginally flat over the AUC is

$$p(d_a, c)=p(AUC, c)\frac{\partial(AUC, c)}{\partial(d_a, c)}=\frac{1}{\tan(2\pi(AUC - 0.5))}\frac{1}{2\sqrt{\pi}}e^{-\frac{d_a^2}{4}}\left[1 - 2\Theta \left(\frac{\frac{c^2-1}{c^2+1}d_a}{\sqrt{2-2\left(\frac{c^2-1}{c^2+1}\right)^2}} \right) \right]. \quad (5)$$

References

1. Pepe, MS. The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford University Press; 2003.
2. Zhou, XH.; Obuchowski, NA.; McClish, DK. Statistical methods in diagnostic medicine. New York, NY: Wiley-Interscience; 2002.
3. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: A tutorial review. Acad Radiol. 2007; 14:723–748. [PubMed: 17502262]
4. Hanley JA. The robustness of the “binormal” assumptions used in fitting ROC curves. Med Decis Making. 1988; 8:197–203. [PubMed: 3398748]
5. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging. 1989; 29:307–335. [PubMed: 2667567]
6. Swets JA. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull. 1986; 99:100–117. [PubMed: 3704032]

7. Swets, JA.; Pickett, RM. Evaluation of diagnostic systems: methods from signal detection theory. New York, NY: Academic Press; 1982.
8. Lloyd CJ. Estimation of a convex ROC curve. *Stat Probabil Lett.* 2002; 59:99–111.
9. Green, DM.; Swets, JA. Signal detection theory and psychophysics. New York, NY: Wiley; 1966.
10. Metz CE. Some practical issues of experimental-design and data-analysis in radiological ROC studies. *Invest Radiol.* 1989; 24:234–245. [PubMed: 2753640]
11. Metz CE, Pan X. “Proper” binormal ROC curves: Theory and maximum-likelihood estimation. *J Math Psychol.* 1999; 43:1–33. [PubMed: 10069933]
12. Pesce LL, Metz CE. Reliable and computationally efficient maximum-likelihood estimation of “proper” binormal ROC curves. *Acad Radiol.* 2007; 14:814–829. [PubMed: 17574132]
13. Zur RM. unpublished data. 2010
14. Gelman, A. Bayesian data analysis. 2nd. Boca Raton, FL: Chapman & Hall/CRC; 2004.
15. Woodworth, GG. Biostatistics: a Bayesian introduction. Hoboken, NJ: Wiley-Interscience; 2005.
16. Shiraishi J, Pesce L, Metz CE, Doi K. On experimental design and data analysis in receiver operating characteristic (ROC) studies: lessons learned from papers published in radiology from 1997 to 2006. *Radiology.* 2009
17. Papoulis, A.; Pillai, SU. Probability, random variables, and stochastic processes. 4th. Boston, MA: McGraw-Hill; 2002.
18. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978; 8:283–298. [PubMed: 112681]
19. Ogilvie JC, Creelman CD. Maximum-likelihood estimation of receiver operating characteristic curve parameters. *J Math Psychol.* 1968; 5:377–391.
20. 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.
21. 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]
22. Kendall, MG.; Stuart, A.; Ord, JK.; Arnold, SF.; O'Hagan, A. Kendall's advanced theory of statistics: classical inference and the linear model. 6th. New York, NY: Hodder Arnold; 1994.
23. Liu, JS. Monte Carlo strategies in scientific computing. New York, NY: Springer; 2001.
24. Robert, CP.; Casella, G. Monte Carlo statistical methods. 2nd. New York, NY: Springer; 2004.
25. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *Am Stat.* 1995; 49:327–335.
26. Neal RM. Slice sampling. *Ann Stat.* 2003; 31:705–741.
27. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol.* 1975; 12:387–415.
28. Scott, DW. Multivariate density estimation: theory, practice, and visualization. New York, NY: Wiley; 1992.
29. Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2nd. Hoboken, NJ: Wiley; 2002.
30. Obuchowski NA, Lieber ML. Confidence bounds when the estimated ROC area is 1.0. *Acad Radiol.* 2002; 9:526–530. [PubMed: 12458878]
31. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Acad Radiol.* 2001; 8:328–334. [PubMed: 11293781]

Figure 1.

Illustration of the flat prior on d_a and c . Panel a) shows AUC values that correspond to combinations of the d_a and c values. The contour lines projected onto the bottom of the figure show isopleths of equal AUC values, ranging from 0.55 to 0.95 in intervals of 0.05. Panel b) shows the probability density distribution of the AUC value given that the probability densities are flat over d_a and c (d_a truncated at 4.0). If the d_a value is not truncated at 4.0, then the probability that the AUC value is approximately 1.0 would have been infinite.

Figure 2. Plot of the log of the flat prior on the AUC in the d_a and c space. The contour lines projected onto the bottom of the figure show isopleths of equal probabilities.

Figure 3.

Example ML and Bayesian ROC curve estimates based on: panels a) and b), a well-behaved dataset, and panels c) and d), a degenerate dataset. The ML estimates of the ROC curves are shown as a heavy solid line and the Bayesian estimates of the ROC curves are shown as 95%, 90%, 75% and 50% iso-probability contours of joint posterior probabilities of TPF and FPF values. Panels a) and c) show Bayesian estimates with the flat prior on d_a and c and panels b) and d) show Bayesian estimates with the flat prior on the AUC. The Bayesian estimates are not shown in panel c) because Bayesian estimation with a flat prior on d_a and c does not converge when estimating ROC curves from degenerate datasets.

Figure 4.

Bias (circles), standard deviation (triangles) and RMSE (solid diamonds) of AUC estimates from ML estimation and from Bayesian MAP and mean-posterior estimations with the flat prior on d_a and c and with the flat prior on the AUC on well-behaved datasets simulated from nine population ROC curves. Uncertainties in the estimates are approximately the size of the data points.

Figure 5.

Comparison of ML and Bayesian estimations in term of the 95% CIs of the AUC estimate from well-behaved datasets simulated from nine population ROC curves. Uncertainties in the coverage are approximately the size of the data points.

Figure 6.

Bias (circles), standard deviation (triangles) and RMSE (solid diamonds) of AUC estimates from ML estimation and from Bayesian MAP and mean-posterior estimations with the flat prior on the AUC on coarsely categorized datasets that are prone to be degenerate, simulated from three population ROC curves and with three different probabilities of producing degenerate datasets. Uncertainties in the estimates are approximately the size of the data points. Bayesian estimates with the flat prior on d_a and c (not shown) do not converge when analyzing degenerate datasets.

Figure 7.

Comparison of ML and Bayesian (with the flat prior on the AUC) estimations in term of the coverage and width of the 95% CIs of the AUC estimates from coarsely categorized datasets that are prone to be degenerate, simulated from three population ROC curves and with three different probabilities of producing degenerate datasets. Uncertainties in the estimates of the coverage are approximately the size of the data points. Bars on the average width represent the range containing 95% of the individual values of the width.

Figure 8.

Comparison between ML and Bayesian mean-posterior (with the flat prior on the AUC) estimates of the AUC (solid lines) and its 95% CIs (points) on coarsely categorized datasets that are prone to be degenerate simulated from three population ROC curves (probability of producing degenerate dataset approximately 50%). The AUC estimates are sorted in increasing order. The dashed lines indicate the AUC values of the population ROC curves.