entropy

MDPI

Article

ROC Analyses Based on Measuring Evidence Using the Relative Belief Ratio

Luai Al-Labadi [1,†], Michael Evans [2,*,†] and Qiaoyu Liang [2,†]

[1] Department of Mathematical and Computational Sciences, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada
[2] Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada
* Correspondence: mevansthree.evans@utoronto.ca
† Authors Al-Labadi, Evans and Liang contributed equally to this work.

Abstract: ROC (Receiver Operating Characteristic) analyses are considered under a variety of assumptions concerning the distributions of a measurement X in two populations. These include the binormal model as well as nonparametric models where little is assumed about the form of distributions. The methodology is based on a characterization of statistical evidence which is dependent on the specification of prior distributions for the unknown population distributions as well as for the relevant prevalence w of the disease in a given population. In all cases, elicitation algorithms are provided to guide the selection of the priors. Inferences are derived for the AUC (Area Under the Curve), the cutoff c used for classification as well as the error characteristics used to assess the quality of the classification.

Keywords: ROC; AUC; optimal cutoff; statistical evidence; relative belief; binormal; mixture Dirichlet process

Citation: Al-Labadi, L.; Evans, M.; Liang, Q. ROC Analyses Based on Measuring Evidence Using the Relative Belief Ratio. *Entropy* **2022**, *24*, 1710. https://doi.org/10.3390/e24121710

Academic Editor: Ciprian Doru Giurcaneanu

Received: 26 September 2022
Accepted: 19 November 2022
Published: 23 November 2022

1. Introduction

An ROC (Receiver Operating Characteristic) analysis is used in medical science to determine whether or not a real-valued diagnostic variable X for a disease or condition is useful. If the diagnostic indicates that an individual has the condition, then this will typically mean that a more expensive or invasive medical procedure is undertaken. So it is important to assess the accuracy of the diagnostic variable X. These methods have a wider class of applications but our terminology will focus on the medical context.

An approach to such analyses is presented here that is based on a characterization of statistical evidence and which incorporates all available information as expressed via prior probability distributions. For example, while p-values are often used in such analyses, there are questions concerning the validity of these quantities as characterizations of statistical evidence. As will be discussed, there are many advantages to the framework adopted here.

A common approach to the assessment of the diagnostic variable X is to estimate its AUC (Area Under the Curve), namely, the probability that an individual sampled from the diseased population will have a higher value of diagnostic variable X than an individual independently sampled from the nondiseased population. A good diagnostic should give a value of the AUC near 1 while a value near $1/2$ indicates a poor diagnostic test (if the AUC is near 0, then the classification is reversed). It is possible, however, that a diagnostic with AUC ≈ 1 may not be suitable (see Examples 1 and 6). In particular, a cutoff value c needs to be selected so that if $X > c$, then an individual is classified as requiring the more invasive procedure. Inferences about the error characteristics for the combination (X, c), such as the false positive rate, etc., are also required.

This paper is concerned with inferences about the AUC, the cutoff c and the error characteristics of the classification based on a valid measure of evidence. A key aspect of the analysis is the *relevant prevalence* w. The phrase "relevant prevalence" means that X will

be applied to a certain population, such as those patients who exhibit certain symptoms, and w represents the proportion of this subpopulation who are diseased. The value of w may vary by geography, medical unit, time, etc. To make a valid assessment of X in an application, it is necessary that the information available concerning w be incorporated. This information is expressed here via an elicited prior probability distribution for w, which may be degenerate at a single value if w is assumed known, or be quite diffuse when little is known about w. In fact, all unknown population quantities are given elicited priors. There are many contexts where data are available relevant to the value of w and this leads to a full posterior analysis for w as well as for the other quantities of interest. Even when such data are not available, however, it is still possible to take the prior for w into account so the uncertainties concerning w always play a role in the analysis and this is a unique aspect of the approach taken here.

While there are some methods available for the choice of c, these often do not depend on the prevalence w which is a key factor in determining the true error characteristics of (X, c) in an application, see [1–5]. So it is preferable to take w into account when considering the value of a diagnostic in a particular context. One approach to choosing c is to minimize some error criterion that depends on w to obtain c_{opt}. As will be demonstrated in the examples, however, sometimes c_{opt} results in a classification that is useless. In such a situation a suboptimal choice of c is required but the error characteristics can still be based on what is known about w so that these are directly relevant to the application.

Others have pointed out deficiencies in the AUC statistic and proposed alternatives. For example, it can be argued that taking into account the costs associated with various misclassification errors is necessary and that using the AUC is implicitly making unrealistic assumptions concerning these costs, see [6]. While costs are relevant, costs are not incorporated here as these are often difficult to quantify. Our goal is to express clearly what the evidence is saying about how good (X, c) is via an assessment of its error characteristics. With the error characteristics in hand, a user can decide whether or not the costs of misclassifications are such that the diagnostic is usable. This may be a qualitative assessment although, if numerical costs are available, these could be subsequently incorporated. The principle here is that economic or social factors be considered separately from what the evidence in the data says, as it is a goal of statistics to clearly state the latter.

The framework for the analysis is Bayesian as proper priors are placed on the unknown distribution F_{ND} (the distribution of X in the nondiseased population), on F_D (the distribution of X in the diseased population) and the prevalence w. In all the problems considered, elicitation algorithms are presented for how to choose these priors. Moreover, all inferences are based on the relative belief characterization of statistical evidence where, for a given quantity, evidence in favor (against) is obtained when posterior beliefs are greater (less) than prior beliefs, see Section 2.2 for discussion and [7]. So evidence is determined by how the data change beliefs. Section 2 discusses the general framework, defines relevant quantities and provides an outline for how specific relative belief inferences are determined. Section 3 develops the inferences for the quantities of interest for three contexts (1) X is an ordered discrete variable with and without constraints on (F_{ND}, F_D) (2) X is a continuous variable and (F_{ND}, F_D) are normal distributions (the *binormal model*) (3) X is a continuous variable and no constraints are placed on (F_{ND}, F_D).

There is previous work on using Bayesian methods in ROC analyses. For example, a Bayesian analysis for the binormal model when there are covariates present is developed in [8]. An estimate of the ROC using the Bayesian bootstrap is discussed in [9]. A Bayesian semiparametric analysis using a Dirichlet mixture process prior is developed in [10,11]. The sampling regime where the data can be used for inference about the relevant prevalence and where a gold standard classifier is not assumed to exist is presented in [12]. Considerable discussion concerning the case where the diagnostic test is binary, covering the cases where there is and is not a gold standard test, as well as the situation where the goal is to compare diagnostic tests and to make inference about the prevalence distribution can be found in [13] and also see [14]. Application of an ROC analysis to a comparison of linear and

nonlinear approaches to a problem in medical physics is in [15]. Further discussion of nonlinear methodology can be found in [16,17].

The contributions of this paper, that have not been covered by previous published work in this area, are as follows:

(i) The primary contribution is to base all the inferences associated with an ROC analysis on a clear and unambiguous characterization of statistical evidence via the principle of evidence and the relative belief ratio. While Bayes factors are also used to measure statistical evidence, there are serious limitations on their usage with continuous parameters as priors are restricted to be of a particular form. The approach via relative belief removes such restrictions on priors and provides a unified treatment of estimation and hypothesis assessment problems. In particular, this leads directly to estimates of all the quantities of interest, together with assessments of the accuracy of the estimates, and a characterization of the evidence, whether in favor of or against a hypothesis, together with a measure of the strength of the evidence. Moreover, no loss functions are required to develop these inferences. The merits of the relative belief approach over others are more fully discussed in Section 2.2.

(ii) A prior on the relevant prevalence is always used to determine inferences even when the posterior distribution of this quantity is not available. As such the prevalence always plays a role in the inferences derived here.

(iii) The error in the estimate of the cut-off is always quantified as well as the errors in the estimates of the characteristics evaluated at the chosen cut-off. It is these characteristics, such as the sensitivity and specificity, that ultimately determine the value of the diagnostic test.

(iv) The hypothesis $H_0 : \text{AUC} > 1/2$ is first assessed and if evidence is found in favor of this, the prior is then conditioned on this event being true for inferences about the remaining quantities. Note that this is equivalent to conditioning the posterior on the event AUC $> 1/2$ when inferences are determined by the posterior but with relative belief inferences both the conditioned prior and conditioned posterior are needed to determine the inferences.

(v) Precise conditions are developed for the existence of an optimal cutoff with the binormal model.

(vi) In the discrete context (1), it is shown how to develop a prior and the analysis under the assumption that the probabilities describing the outcomes from the diagnostic variable X are monotone.

The relative belief ratio, as a measure of evidence, is seen to have a connection to relative entropy. For example, it is equivalent, in the sense that the inferences are the same, to use the logarithm of the relative belief ratio as the measure of evidence. The relative entropy is then the posterior expectation of this quantity and so can be considered as a measure of the overall evidence provided by the model, prior and data concerning a quantity of interest.

The methods used for all the computations in the paper are simulation based and represent fairly standard Bayesian computational methods. In each context considered, sufficient detail is provided so that these can be implemented by a user.

2. The Problem

Consider the formulation of the problem as presented in [18,19] but with somewhat different notation. There is a measurement $X : \Omega \to R^1$ defined on a population $\Omega = \Omega_D \cup \Omega_{ND}$, with $\Omega_D \cap \Omega_{ND} = \phi$, where Ω_D is comprised of those with a particular disease, and Ω_{ND} represents those without the disease. So $F_{ND}(c) = \#(\{\omega \in \Omega_{ND} : X(\omega) \leq c\})/\#(\Omega_{ND})$ is the conditional cdf of X in the nondiseased population, and $F_D(x) = \#(\{\omega \in \Omega_D : X(\omega) \leq x\})/\#(\Omega_D)$ is the conditional cdf of X in the diseased population. It is assumed that there is a gold standard classifier, typically much more difficult to use than X, such that for any $\omega \in \Omega$ it can be determined definitively if $\omega \in \Omega_D$ or $\omega \in \Omega_{ND}$. There are two ways in which one can sample from Ω, namely,

(i) take samples from each of Ω_D and Ω_{ND} separately or
(ii) take a sample from Ω.

The sampling method used affects the inferences that can be drawn. For many studies (i) is the relevant sampling mode, as in case-control studies, while (ii) is relevant in cross-sectional studies.

It supposed that the greater the value $X(\omega)$ is for individual ω, the more likely it is that $\omega \in \Omega_D$. For the classification, a cutoff value c is required such that, if $X(\omega) > c$, then ω is classified as being in Ω_D and otherwise is classified as being in Ω_{ND}. However, X is an imperfect classifier for any c and it is necessary to assess the performance of (X, c). It seems natural that a value of c be used that is optimal in some sense related to the error characteristics of this classification. Table 1 gives the relevant probabilities for classification into Ω_D and Ω_{ND}, together with some common terminology, in a *confusion matrix*.

Table 1. Error probabilities when $X > c$ indicates a positive.

	$\boldsymbol{\Omega_D}$	$\boldsymbol{\Omega_{ND}}$
$X > c$	$\mathrm{TPR}(c) = 1 - F_D(c)$ *sensitivity* (*recall*) or true positive rate	$\mathrm{FPR}(c) = 1 - F_{ND}(c)$ false positive rate
$X \leq c$	$\mathrm{FNR}(c) = F_D(c)$ false negative rate	$\mathrm{TNR}(c) = F_{ND}(c)$ *specificity* or true negative rate

Another key ingredient is the prevalence $w = \#(\Omega_D)/\#(\Omega)$ of the disease in Ω. In practical situations, it is necessary to also take w into account in assessing the error in (X, c). The following error characteristics depend on w,

$$\begin{aligned}
\mathrm{Error}(c) &= \text{misclassification rate} = w\mathrm{FNR}(c) + (1-w)\mathrm{FPR}(c),\\
\mathrm{FDR}(c) &= \text{false discovery rate} = \frac{(1-w)\mathrm{FPR}(c)}{w(1-\mathrm{FNR}(c)) + (1-w)\mathrm{FPR}(c)},\\
\mathrm{FNDR}(c) &= \text{false nondiscovery rate} = \frac{w\mathrm{FNR}(c)}{w\mathrm{FNR}(c) + (1-w)(1-\mathrm{FPR}(c))}.
\end{aligned}$$

Under sampling regime (ii) and cutoff c, $\mathrm{Error}(c)$ is the probability of making an error, $\mathrm{FDR}(c)$ is the conditional probability of a subject being misclassified as positive given that it has been classified as positive and $\mathrm{FNDR}(c)$ is the conditional probability of a subject being misclassified as negative given that it has been classified as negative. In other words, $\mathrm{FDR}(c)$ is the proportion of those individuals in the population consisting of those who have been classified by the diagnostic test as having the disease, but in fact do not have it. It is often observed that when w is very small and $\mathrm{FNR}(c)$ and $\mathrm{FPR}(c)$ are small, then $\mathrm{FDR}(c)$ can be big. This is sometimes referred to as the *base rate fallacy* as, even though the test appears to be a good one, there is a high probability that an individual classified as having the disease will be misclassified. For example, if $w = \mathrm{FNR}(c) = \mathrm{FPR}(c) = 0.05$, then $\mathrm{Error}(c) = 0.05$, $\mathrm{FDR}(c) = 0.50$, $\mathrm{FNDR}(c) = 2.76 \times 10^{-3}$ and when $w = 0.01$, then $\mathrm{Error}(c) = 0.05$, $\mathrm{FDR}(c) = 0.84$, $\mathrm{FNDR}(c) = 5.31 \times 10^{-4}$. In these cases the false nondiscovery rate is quite small while the false discovery rate is large. If the disease is highly contagious, then these probabilities may be considered acceptable but indeed they need to be estimated. Similarly, $\mathrm{FNDR}(c)$ may be small when $\mathrm{FNR}(c)$ is large and w is very small.

It is naturally desirable to make inference about an optimal cutoff c_{opt} and its associated error quantities. For a given value of w, the optimal cutoff will be defined here as $c_{opt} = \arg\inf \mathrm{Error}(c)$, the value which minimizes the probability of making an error. Other choices for determining a c_{opt} can be made, and the analysis and computations will be similar, but our thesis is that, when possible, any such criterion should involve the prior

distribution of the relevant prevalence w. As demonstrated in Example 6 this can sometimes lead to useless values of c_{opt} even when the AUC is large. While this situation calls into question the value of the diagnostic, a suboptimal choice of c can still be made according to some alternative methodology. For example, sometimes *Youden's index*, which maximizes $1 - 2\text{Error}(c)$ over c with $w = 1/2$, is recommended, or the *closest-to-(0,1)* criterion which minimizes $\text{FPR}(c)^2 + (1 - \text{TPR}(c))^2$, see [2] for discussion. Youden's index and the closest-to-(0,1) criterion do not depend on the prevalence and have geometrical interpretations in terms of the ROC curve, but as we will see, the ROC curve does not exist in full generality and this is particularly relevant in the discrete case. The methodology developed here provides an estimate of the c to be used, together with an exact assessment of the error in this estimate, as well as providing estimates of the associated error characteristics of the classification.

Letting $\hat{c}_{opt}$ denote the estimate of c_{opt}, the values of $\text{Error}(\hat{c}_{opt})$, $\text{TPR}(\hat{c}_{opt})$, $\text{FPR}(\hat{c}_{opt})$, $\text{FNR}(\hat{c}_{opt})$ and $\text{TNR}(\hat{c}_{opt})$ are also estimated and the recorded values used to assess the value of the diagnostic test. There are also other characteristics that may prove useful in this regard such as the *positive predictive value* (PPV)

$$\text{PPV}(c) = \frac{w\text{TPR}(c)}{w\text{TPR}(c) + (1-w)\text{FPR}(c)},$$

namely, the conditional probability a subject is positive given that they have tested positive, which plays a role similar to $\text{FDR}(c)$. See [14] for discussion of the PPV and the similarly defined *negative predictive value* (NPV). The value of $\text{PPV}(\hat{c}_{opt})$ can be estimated in the same way as the other quantities as is subsequently discussed.

2.1. The AUC and ROC

Consider two situations where F_{ND}, F_D are either both absolutely continuous or both discrete. In the discrete case, suppose that these distributions are concentrated on a set of points $c_1 < c_2 < \cdots < c_m$. When ω_D, ω_{ND} are selected using sampling scheme (i), then the probability that a higher score is received on diagnostic X by a diseased individual than a nondiseased individual is

$$\text{AUC} = \begin{cases} \int_{-\infty}^{\infty} (1 - F_D(c))\, f_{ND}(c)\, dc & \text{abs. cont.} \\ \sum_{i=1}^{m} (1 - F_D(c_i))(F_{ND}(c_i) - F_{ND}(c_{i-1})) & \text{discrete.} \end{cases} \tag{1}$$

Under the assumption that $F_D(c)$ is constant on $\{c : F_{ND}(c) = p\}$ for every $p \in [0,1]$, there is a function ROC (*receiver operating characteristic*) such that $1 - F_D(c) = \text{ROC}(1 - F_{ND}(c))$ so $\text{AUC} = \int_{-\infty}^{\infty} \text{ROC}(1 - F_{ND}(c))\, F_{ND}(dx)$. Putting $p = 1 - F_{ND}(c)$, then $\text{ROC}(p) = 1 - F_D(F_{ND}^{-1}(1-p))$. In the absolutely continuous case, $\text{AUC} = \int_0^1 \text{ROC}(p)\, dp$ which is the *area under the curve* given by the ROC function. The area under the curve interpretation is geometrically evocative but is not necessary for (1) to be meaningful.

It is commonly suggested that a good diagnostic variable X will have an AUC close to 1 while a value close to $1/2$ suggests a poor diagnostic test. It is surely the case, however, that the utility of X in practice will depend on the cutoff c chosen and the various error characteristics associated with this choice. So while the AUC can be used to screen diagnostics, it is only part of the analysis and inferences about the error characteristics are required to truly assess the performance of a diagnostic. Consider an example.

Example 1. *Suppose that $F_D = F_{ND}^q$ for some $q > 1$, where F_{ND} is continuous, strictly increasing with associated density f_{ND}. Then using (1), $\text{AUC} = 1 - 1/(q+1)$ which is approximately 1 when q is large. The optimal c minimizes $\text{Error}(c) = wF_{ND}^q(c) + (1-w)(1 - F_{ND}(c))$ which implies c satisfies $F_{ND}(c) = \{(1-w)/qw\}^{1/(q-1)}$ when $q > (1-w)/w$ and the optimal c is otherwise $c = \infty$. If $q = 99$, then $\text{AUC} = 0.99$ and with $w = 0.025$, $(1-w)/w = 39 < q$ so $\text{FNR}(c_{opt}) = 0.390$, $\text{FPR}(c_{opt}) = 0.009$, $\text{Error}(c_{opt}) = 0.019$, $\text{FDR}(c_{opt}) = 0.009$ and*

$FNDR(c_{opt}) = 0.010$. So X seems like a good diagnostic via the AUC and the error characteristics that depend on the prevalence although within the diseased population the probability is 0.39 *of not detecting the disease. If instead $w = 0.01$, then the AUC is the same but $q = 99 = (1-w)/w$ and the optimal classification always classifies an individual as non-diseased which is useless. So the AUC does not indicate enough about the characteristics of the diagnostic to determine if it is useful or not. It is necessary to look at the error characteristics of the classification at the cutoff value that will actually be used, to determine if a diagnostic is suitable and this implies that information about w is necessary in an application.*

2.2. Relative Belief Inferences

Suppose there is a model $\{f_\theta : \theta \in \Theta\}$ for data x together with a prior probability measure Π, with density π, on Θ. These ingredients lead, via the *principle of conditional probability*, to beliefs about the true value of θ, as initially expressed by Π, being replaced by the posterior probability measure $\Pi(\cdot \mid x)$ with density $\pi(\cdot \mid x)$. Note that if interest is instead in a quantity $\psi = \Psi(\theta)$, where $\Psi : \Theta \to \Psi$ and we use the same notation for the function and its range, then the model is replaced by $\{m_\psi : \psi \in \Psi\}$, where $m_\psi(x) = \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\pi(\theta \mid \psi)\, d\theta$ is obtained by integrating out the nuisance parameters, and the prior is replaced by the marginal prior $\pi_\Psi(\psi) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta)\, d\theta$. This leads to the marginal posterior $\Pi_\Psi(\cdot \mid x)$ with density $\pi_\Psi(\cdot \mid x)$.

For the moment suppose that all the distributions are discrete. The *principle of evidence* then says that there is evidence in favor of the value ψ if $\pi_\Psi(\psi \mid x) > \pi_\Psi(\psi)$, evidence against the value ψ if $\pi_\Psi(\psi \mid x) < \pi_\Psi(\psi)$, and no evidence either way if $\pi_\Psi(\psi \mid x) = \pi_\Psi(\psi)$. So, for example, there is evidence in favor of ψ if the probability of ψ increases after seeing the data. To order the possible values with respect to the evidence, we use the *relative belief ratio*

$$RB_\Psi(\psi \mid x) = \frac{\pi_\Psi(\psi \mid x)}{\pi_\Psi(\psi)}.$$

Note that $RB_\Psi(\psi \mid x) > (<) 1$ indicates whether there is evidence in favor of (against) the value ψ. If there is evidence in favor of both ψ_1 and ψ_2, then there is more evidence in favor of ψ_1 than ψ_2 whenever $RB_\Psi(\psi_1 \mid x) > RB_\Psi(\psi_2 \mid x)$ and, if there is evidence against both ψ_1 and ψ_2, then there is more evidence against ψ_1 than ψ_2 whenever $RB_\Psi(\psi_1 \mid x) < RB_\Psi(\psi_2 \mid x)$. For the continuous case consider a sequence of neighborhoods $N_\epsilon(\psi) \downarrow \{\psi\}$ as $\epsilon \to 0$ and then

$$RB_\Psi(N_\epsilon(\psi) \mid x) = \frac{\Pi_\Psi(N_\epsilon(\psi) \mid x)}{\Pi_\Psi(N_\epsilon(\psi))} \to \frac{\pi_\Psi(\psi \mid x)}{\pi_\Psi(\psi)} \tag{2}$$

under very weak conditions such as $\pi_\Psi(\psi) > 0$ and π_Ψ being continuous at ψ.

All the inferences about quantities considered in the paper are derived based upon the principle of evidence as expressed via the relative belief ratio. For example, it is immediate that the value $RB_\Psi(\psi_0 \mid x)$ indicates whether or not there is evidence in favor of or against the hypothesis $H_0 : \Psi(\theta) = \psi_0$. Furthermore, the posterior probability $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$ measures the strength of this evidence for, if $RB_\Psi(\psi_0 \mid x) > 1$ and this probability is large, then there is strong evidence in favor of H_0 as there is a small belief that the true value has a larger relative belief ratio and if $RB_\Psi(\psi_0 \mid x) < 1$ and this probability is small, then there is strong evidence against H_0 as there is high belief that the true value has a larger relative belief ratio. For estimation it is natural to estimate ψ by the *relative belief estimate* $\psi(x) = \arg\sup_{\psi \in \Psi} RB_\Psi(\psi \mid x)$ as this value has the maximum evidence in its favor. Furthermore, the accuracy of this estimate can be assessed by looking at the *plausible region* $Pl_\Psi(x) = \{\psi : RB_\Psi(\psi \mid x) > 1\}$, consisting of all those values for which there is evidence in favor, together with its size and posterior content which measures how strongly it is believed the true value lies in this set. Rather than using the plausible region to assess the accuracy of $\psi(x)$, one could quote a $\gamma-$*relative belief credible region*

$$C_{\Psi,\gamma}(x) = \{\psi : RB_\Psi(\psi \mid x) > c_\gamma\}$$

where the constant c_γ is the largest value such that $\Pi_\Psi(C_{\Psi,\gamma}(x) \mid x) \geq \gamma$. It is necessary, however, that $\gamma \leq \Pi_\Psi(Pl_\Psi(x) \mid x)$ as otherwise $C_{\Psi,\gamma}(x)$ will contain values for which there is evidence against, and this is only known after the data have been seen.

It is established in [7], and in papers referenced there, that these inferences possess a number of good properties such as consistency, satisfy various optimality criteria and clearly they are based on a direct measure of the evidence. Perhaps most significant is the fact that all the inferences are invariant under reparameterizations. For if $\lambda = \Lambda(\psi)$, where Λ is a smooth bijection, then

$$RB_\Lambda(\lambda \mid x) = \frac{\pi_\Lambda(\lambda \mid x)}{\pi_\Lambda(\lambda)} = \frac{\pi_\Psi(\Lambda^{-1}(\lambda) \mid x) J_\Lambda(\Lambda^{-1}(\lambda))}{\pi_\Psi(\Lambda^{-1}(\lambda)) J_\Lambda(\Lambda^{-1}(\lambda))} = \frac{\pi_\Psi((\psi) \mid x)}{\pi_\Psi(\psi)}$$

and so, for example, $\lambda(x) = \Lambda(\psi(x))$. This invariance property is not possessed by the most common inference methods employed such as MAP estimation or using posterior means and this invariance holds no matter what the dimension of ψ is. Moreover, it is proved in [20] that relative belief inferences are optimally robust among all Bayesian inferences for ψ, to linear contaminations of the prior on ψ.

An analysis, using relative belief, of the data obtained in several physics experiments that were all concerned with examining whether there was evidence in favor of or against the quantum model versus hidden variables is available in [21]. Furthermore, an approach to checking models used for quantum mechanics via relative belief is discussed in [22]. Other applications of relative belief inferences to common problems of statistical practice can be found in [7].

The Bayes factor is an alternative measure of evidence and is commonly used for hypothesis assessment in Bayesian inference. To see why the relative belief ratio has advantages over the Bayes factor for evidence-based inferences consider first assessing the hypothesis $H_0 : \Psi(\theta) = \psi_0$. When the prior probability of ψ_0 satisfies $0 < \Pi_\Psi(\{\psi_0\}) < 1$, then the Bayes factor is defined as the ratio of the posterior odds in favor of H_0 to the prior odds in favor of H_0, namely,

$$BF_\Psi(\psi_0 \mid x) = \left\{ \frac{\Pi_\Psi(\{\psi_0\} \mid x)}{\Pi_\Psi(\{\psi_0\}^c \mid x)} \right\} \left\{ \frac{\Pi_\Psi(\{\psi_0\})}{\Pi_\Psi(\{\psi_0\}^c)} \right\}^{-1}.$$

It is easily shown that the Bayes factor satisfies the principle of evidence and $BF_\Psi(\psi_0 \mid x) > (<) 1$ is evidence in favor (against) H_0, so in this context it is a valid measure of evidence.

One might wonder why it is necessary to consider a ratio of odds as opposed to the simpler ratio of probabilities, as specified by the relative belief ratio, for the purpose of measuring evidence but in fact there is a more serious issue with the Bayes factor. For suppose, as commonly arises in applications, that Π_Ψ is a continuous probability measure so that $\Pi_\Psi(\{\psi_0\}) = 0$ as then the Bayes factor for H_0 is not defined. The common recommendation in this context is to require the specification of the following ingredients: a prior probability $p > 0$, a prior distribution Π_{H_0} concentrated on $\Psi^{-1}\{\psi_0\}$ which provides the prior predictive density $m_{H_0}(x)$, a prior distribution $\Pi_{H_0^c}$ concentrated on $\Psi^{-1}\{\psi_0\}^c$ which provides the prior predictive density $m_{H_0^c}(x)$ and then the full prior is taken to be the mixture $\Pi = p\Pi_{H_0} + (1-p)\Pi_{H_0^c}$. With this prior the Bayes factor for H_0 is defined, as now the prior probability of ψ_0 equals p, and an easy calculation shows that $BF_\Psi(\psi_0 \mid x) = m_{H_0}(x)/m_{H_0^c}(x)$. Typically the prior $\Pi_{H_0^c}$ is taken to be the prior that we might place on θ when interest is in estimating ψ.

Now consider the problem of estimating ψ and the prior is such that $\Pi_\Psi(\{\psi\}) = 0$ for every value of ψ as with a continuous prior. The Bayes factor is then not defined for any value of ψ and, if we wished to use the Bayes factor for estimation purposes, it would be necessary to modify the prior to be a different mixture for each value of ψ so that

there would be in effect multiple different priors. This does not correspond to the logic underlying Bayesian inference. When using the relative belief ratio for inference only one prior is required and the same measure of evidence is used for both hypothesis assessment and estimation purposes.

Another approach to dealing with the problem that arises with the Bayes factor and continuous priors is to take a limit as in (2) and, when this is done, we obtain the result

$$BF_{\Psi}(N_{\epsilon}(\psi) \mid x) \to RB_{\Psi}(\psi \mid x)$$

as $\epsilon \to 0$ whenever the prior density of Ψ is continuous and positive at ψ. In other words the relative belief ratio can be also considered as a natural definition of the Bayes factor in continuous contexts.

3. Inferences for an ROC Analysis

Suppose we have a sample of n_D from Ω_D, namely, $x_D = (x_{D1}, \ldots, x_{Dn_D})$ and a sample of n_{ND} from Ω_{ND}, namely, $x_{ND} = (x_{ND1}, \ldots, x_{NDn_{ND}})$ and the goal is to make inference about the AUC, the cutoff c and the error characteristics $\text{FNR}(c)$, $\text{FPR}(c)$, $\text{Error}(c)$, $\text{FDR}(c)$ and $\text{FNDR}(c)$. For the AUC it makes sense to first assess the hypothesis $H_0 : \text{AUC} > 1/2$ via stating whether there is evidence for or against H_0 together with an assessment of the strength of this evidence. Estimates are required for all of these quantities, together with an assessment of the accuracy of the estimate.

3.1. The Prevalence

Consider first inferences for the relevant prevalence w. If w is known, or at least assumed known, then nothing further needs to be done but otherwise this quantity needs to be estimated when assessing the value of the diagnostic and so uncertainty about w needs to be addressed.

If the full data set is based on sampling scheme (ii), then $n_D \sim \text{binomial}(n, w)$. A natural prior π_W to place on w is a $\text{beta}(\alpha_{1w}, \alpha_{2w})$ distribution. The hyperparameters are chosen based on the elicitation algorithm discussed in [23] where interval $[l, u]$ is chosen such that it is believed that $w \in [l, u]$ with prior probability γ. Here $[l, u]$ is chosen so that we are virtually certain that $w \in [l, u]$ and $\gamma = 0.99$ then seems like a reasonable choice. Note that choosing $l = u$ corresponds to w being known and so $\gamma = 1$ in that case. Next pick a point $\xi_w \in [l, u]$ for the mode of the prior and a reasonable choice might be $\xi_w = (l + u)/2$. Then putting $\tau_w = \alpha_{1w} + \alpha_{2w} - 2$ leads to the parameterization $\text{beta}(\alpha_{1w}, \alpha_{2w}) = \text{beta}(1 + \tau_w \xi_w, 1 + \tau_w(1 - \xi_w))$ where ξ_w locates the mode and τ_w controls the spread of the distribution about ξ_w. Here $\tau_w = 0$ gives the uniform distribution and $\tau_w = \infty$ gives the distribution degenerate at ξ_w. With ξ_w specified, τ_w is the smallest value of τ_w such that the probability content of $[l, u]$ is γ and this is found iteratively. For example, if $[l, u] = [0.60, 0.70]$ and $\gamma = 0.99$, so w is known reasonably well, then $\xi_w = (l + u)/2 = 0.65$ and $\tau_w = 601.1$, so the prior is $\text{beta}(391.72, 211.39)$ and the posterior is $\text{beta}(391.72 + n_D, 211.39 + n_{ND})$.

The estimate of w is then

$$w(n_D, n_{ND}) = \arg \sup_{w \in [0,1]} RB(w \mid n_D, n_{ND}) = \arg \sup_{w \in [0,1]} \frac{\pi_W(w \mid n_D, n_{ND})}{\pi_W(w)}.$$

In this case the estimate is the MLE, namely, $w(n_D, n_{ND}) = n_D/(n_D + n_{ND})$. The accuracy of this estimate is measured by the size of the plausible region $Pl(n_D, n_{ND}) = \{w : RB(w \mid n_D, n_{ND}) > 1\}$. For example, if $n = 100$ and $n_D = 68$, then $w(68, 32) = 0.68$ and $Pl(68, 32) = [0.647, 0.712]$ which has posterior content 0.651. So the data suggest that the upper bound of $u = 0.70$ is too strong although the posterior belief in this interval is not very high.

The prior and posterior distributions of w play a role in inferences about all the quantities that depend on the prevalence. In the case where the cutoff is determined by minimizing the probability of a misclassification, then c_{opt}, $\text{FNR}(c_{opt})$, $\text{FPR}(c_{opt})$, $\text{Error}(c_{opt})$, $\text{FDR}(c_{opt})$ and $\text{FNDR}(c_{opt})$ all depend on the prevalence. Under sampling scheme (i), however, only the prior on w has any influence when considering the effectiveness of X. Inference for these quantities is now discussed in both cases.

3.2. Ordered Discrete Diagnostic

Suppose X takes values on the finite ordered scale $c_1 < c_2 < \cdots < c_m$ and let $p_{NDi} = P(X(\omega_{ND}) = c_i)$, $p_{Di} = P(X(\omega_D) = c_i)$ so $F_{ND}(c_i) = \sum_{j=1}^{i} p_{NDj}$ and $F_D(c_i) = \sum_{j=1}^{i} p_{Dj}$. These imply that $\text{FPR}(c_i) = 1 - \sum_{j=1}^{i} p_{NDi}$, $\text{FNR}(c_i) = \sum_{j=1}^{i} p_{Di}$,

$$\text{AUC}(p_{ND}, p_D) = \sum_{i=1}^{m} (1 - \text{FNR}(c_i)) p_{NDi}$$

with the remaining quantities defined similarly. Ref. [23] can be used to obtain independent elicited Dirichlet priors

$$p_{ND} \sim \text{Dirichlet}(\alpha_{ND1}, \ldots, \alpha_{NDm}),\ p_D \sim \text{Dirichlet}(\alpha_{D1}, \ldots, \alpha_{Dm}) \tag{3}$$

on these probabilities by placing either upper or lower bounds on each cell probability that hold with virtual certainty γ, as discussed for the beta prior on the prevalence. If little information is available, it is reasonable to use uniform ($\text{Dirichlet}(1, \ldots, 1)$) priors on p_{ND} and p_D. This together with the independent prior on w leads to prior distributions for the AUC, c_{opt} and all the quantities associated with error assessment such as $\text{FNR}(c_{opt})$, etc.

Data (x_D, x_{ND}) lead to counts $f_{ND} = (f_{ND1}, \ldots, f_{NDm})$ and $f_D = (f_{D1}, \ldots, f_{Dm})$ which in turn lead to the independent posteriors

$$p_{ND} \mid f_{ND} \sim \text{Dirichlet}(\alpha_{ND} + f_{ND}),\ p_D \mid f_D \sim \text{Dirichlet}(\alpha_D + f_D). \tag{4}$$

Under sampling regime (ii) this, together with the independent posterior on w, leads to posterior distributions for all the quantities of interest. Under sampling regime (i), however, the logical thing to do, so the inferences reflect the uncertainty about w, is to only use the prior on w when deriving inferences about any quantities that depend on this such as c_{opt} and the various error assessments.

Consider inferences for the AUC. The first inference should be to assess the hypothesis $H_0 : \text{AUC} > 1/2$ for, if H_0 is false, then X would seem to have no value as a diagnostic (the possibility that the directionality is wrong is ignored here). The relative belief ratio $RB(H_0 \mid f_{ND}, f_{ND}) = \Pi(H_0 \mid f_{ND}, f_{ND})/\Pi(H_0)$ is computed and compared to 1. If it is concluded that H_0 is true, then perhaps the next inference of interest is to estimate the AUC via the relative belief estimate. The prior and posterior densities of the AUC are not available in closed form so estimates are required and density histograms are employed here for this. The set $(0, 1]$ is discretized into L subintervals $(0, 1] = \cup_{i=1}^{L}((i-1)/L, i/L]$, and putting $a_i = (i - 1/2)/L$, the value of the prior density $p_{\text{AUC}}(a_i)$ is estimated by L(proportion of prior simulated values of AUC in $(i-1, i]/L$) and similarly for the posterior density $p_{\text{AUC}}(a_i \mid f_{ND}, f_D)$. Then $RB_{\text{AUC}}(a \mid f_{ND}, f_{ND})$ is maximized to obtain the relative belief estimate $\text{AUC}(f_{ND}, f_D)$ together with the plausible region and its posterior content.

These quantities are also obtained for c_{opt} in a similar fashion, although c_{opt} has prior and posterior distribution concentrated on $\{c_1, c_2, \ldots, c_m\}$ so there is no need to discretize. Estimates of the quantities $\text{FNR}(c_{opt}(f_{ND}, f_D))$, $\text{FPR}(c_{opt}(f_{ND}, f_D))$, $\text{Error}(c_{opt}(f_{ND}, f_D))$, $\text{FDR}(c_{opt}(f_{ND}, f_D))$ and $\text{FNDR}(c_{opt}(f_{ND}, f_D))$ are also obtained as these indicate the performance of the diagnostic in practice. The relative belief estimates of these quantities are easily obtained in a second simulation where $c_{opt}(f_{ND}, f_D)$ is fixed.

Consider now an example.

Example 2. *Simulated example.*

For $k = 5$ and $c_i = i$, data were generated as

$$f_{ND} \sim multinomial(50, 0.5, 0.2, 0.1, 0.1, 0.1) \text{ obtaining } f_{ND} = (29, 7, 4, 5, 5),$$
$$f_D \sim multinomial(100, 0.1, 0.1, 0.2, 0.3, 0.3) \text{ obtaining } f_D = (14, 7, 25, 33, 21).$$

With these choices for p_{ND}, p_D the true values are AUC$= 0.65$, and with $w = 0.65$, $c_{opt} = 2$, $FNR(c_{opt}) = 0.200$, $FPR(c_{opt}) = 0.300$, $Error_w(c_{opt}) = 0.235$, $FDR(c_{opt}) = 0.168$ and $FNDR(c_{opt}) = 0.347$. So X is not an outstanding diagnostic but with these error characteristics it may prove suitable for a given application. Uniform, namely, Dirichlet$(1, 1, 1, 1, 1)$, priors were placed on p_{ND} and p_D, reflecting little knowledge about these quantities.

Simulations based on Monte Carlo sample sizes of $N = 10^5$ from the prior and posterior distributions of p_{ND} and p_D were conducted and the prior and posterior distributions of the quantities of interest obtained. The hypothesis $H_0 : AUC > 0.5$ is assessed by $RB_{AUC}((0.50, 1.00] \mid f_{ND}, f_D) = 3.15$. So there is evidence in favor of H_0 and the strength of this evidence is measured by the posterior probability content of $(0.50, 1.00]$ which equals 1.0 to machine accuracy and so this is categorical evidence in favor of H_0.

For the continuous quantities a grid based on $L + 1 = 25$ equispaced points $\{0, 0.04, 0.08, \ldots, 1.00\}$ was used and all the mass in the interval $(i - 1, i]/L$ assigned to the midpoint $(i - 1/2)/L$. Figure 1 contains plots of the prior and posterior densities and relative belief ratio of the AUC. The relative belief estimate of the AUC is $AUC(f_{ND}, f_D) = 0.66$ with $Pl_{AUC}(f_{ND}, f_D) = [0.60, 0.72]$ having posterior content 0.97. Certainly a finer partition of $[0, 1]$ than just 24 intervals is possible, but even in this relatively coarse case the results are quite accurate.

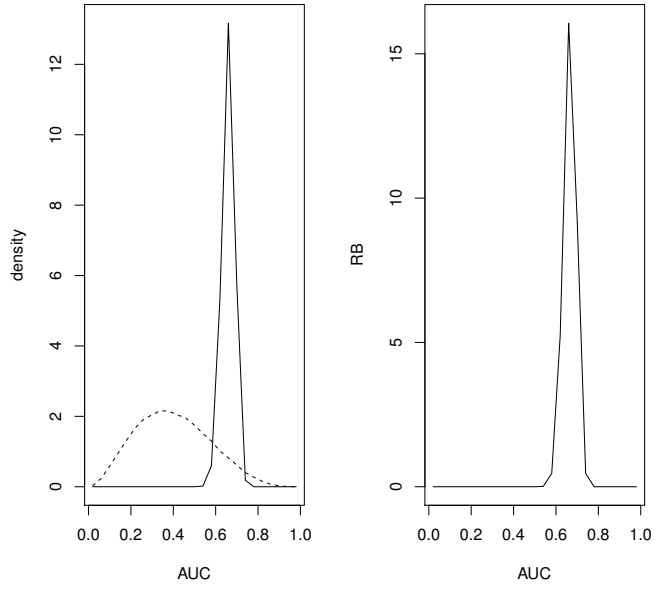

Figure 1. In Example 2, plots of the prior (- - -), the posterior (—) and the RB ratio of the AUC.

Supposing that the relevant prevalence is known to be $w = 0.65$, Figure 2 contains plots of the prior and posterior densities and relative belief ratio of c_{opt}. The relative belief estimate is $c_{opt}(f_{ND}, f_D) = 2$ with $Pl_{c_{opt}}(f_{ND}, f_D) = \{2\}$ with posterior probability content 0.53 so the correct optimal cut-off has been identified but there is a degree of uncertainty concerning this. The error characteristics that tell us about the utility of X as a diagnostic are given by the relative belief estimates (column (a)) in Table 2. It is interesting to note that the estimate of $Error(c_{opt})$ is determined by the prior and posterior distributions of a convex combination of $FPR(c_{opt})$ and $FNR(c_{opt})$ and the estimate is not the same convex combination of the estimates of $FPR(c_{opt})$ and $FNR(c_{opt})$. So, in this case $Error(c_{opt})$ seems like a much better assessment of the performance of the diagnostic.

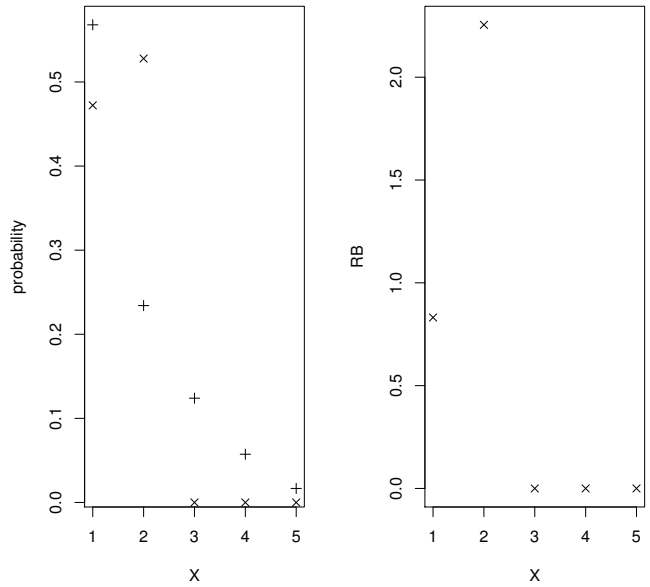

Figure 2. In Example 2, plots of the the prior (+), the posterior (×) and the RB ratio of c_{opt}.

Suppose now that the prevalence is not known but there is a beta$(1+\tau_w\xi_w, 1+\tau_w(1-\xi_w))$ *prior specified for* w *and consider the choice discussed in Section 3.1 where* $\xi_w = 0.65$ *and* $\tau_w = 601.1$. *When the data are produced according to sampling regime (i), then there is no posterior for* w *but this prior can still be used in determining the prior and posterior distributions of* c_{opt} *and the associated error characteristics. When this simulation was carried out* $c_{opt}(f_{ND}, f_D) = 2$ *with* $Pl_{c_{opt}}(f_{ND}, f_D) = \{2\}$ *with posterior probability content* 0.53. *and column (b) of Table 2 gives the estimates of the error characteristics. So other than the estimate of the FPR, the results are similar. Finally, assuming that the data arose under sampling scheme (ii), then* w *has a posterior distribution and using this gives* $c_{opt}(f_{ND}, f_D) = 2$ *with* $Pl_{c_{opt}}(f_{ND}, f_D) = \{2\}$ *with posterior probability content* 0.52 *and error characteristics as in column (c) of Table 2. These results are the same as if the prevalence is known which is sensible as the posterior concentrates about the true value more than the prior.*

Table 2. The estimates of the error characteristcs of X at $c_{opt} = 2$ in Example 2 where (a) w is assumed known, (b) only the prior for w is available, (c) the posterior for w is also available.

Quantity	Estimate (a)	Estimate (b)	Estimate (c)
FPR(c_{opt})	0.30	0.26	0.30
FNR(c_{opt})	0.22	0.22	0.22
Error(c_{opt})	0.22	0.22	0.22
FDR(c_{opt})	0.14	0.14	0.14
FNDR(c_{opt})	0.34	0.34	0.34

Another somewhat anomalous feature of this example is the fact that uniform priors on p_D *and* p_{ND} *do not lead to a prior on the AUC that is even close to uniform. In fact one could say that this prior has a built-in bias against a diagnostic with AUC* $> 1/2$ *and indeed most choices of* p_D *and* p_{ND} *will not satisfy this. Another possibility is to require* $p_{ND1} \geq \cdots \geq p_{NDm}$ *and* $p_{D1} \leq \cdots \leq p_{Dm}$, *namely, require monotonicity of the probabilities. A result in [22] implies that* p_{ND} *satisfies this iff* $p_{ND} = A_k\omega_{ND}$ *where* $\omega_{ND} \in S_k$, *the standard* $(k-1)$*-dimensional simplex, and* $A_k \in R^{k\times k}$ *with i-ith row equal to* $(0,\ldots,0,1/i,1/(i+1),\ldots,1/k)$ *and* p_D *satisfies this iff* $p_D = B_k\omega_D$ *where* $\omega_D \in S_k$ *and* $B_k = I_k^* A_k$ *where* $I_k^* \in R^{k\times k}$ *contains all 0's except for 1's on the crossdiagonal. If* ω_{ND} *and* ω_D *are independent and uniform on* S_k, *then* p_D *and* p_{ND} *are*

independent and uniform on the sets of probabilities satisfying the corresponding monotonicities and Figure 3 has a plot of the prior of the AUC when this is the case. It is seen that this prior is biased in favor of $AUC > 1/2$. *Figure 3 also has a plot of the prior of the AUC when* p_D *is uniform on the set of all nondecreasing probabilities and* p_{ND} *is uniform on* S_k. *This reflects a much more modest belief that X will satisfy* $AUC > 1/2$ *and indeed this may be a more appropriate prior than using uniform distributions on* S_k. *Ref. [22] also provides elicitation algorithms for choosing alternative Dirichlet distributions for* ω_{ND} *and* ω_D.

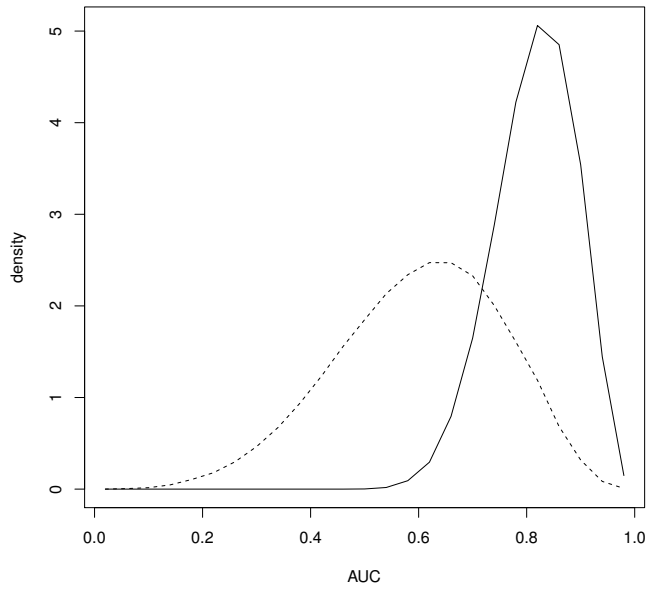

Figure 3. Prior density of the AUC when p_D is uniform on the set of nondecreasing probabilities independent of p_{ND} uniform on the set of nonincreasing probabilities (–) as well as when p_D is uniformly distributed on the set of nondecreasing probabilities independent of p_{ND} uniform on S_k (- -).

When $H_0 : AUC > 0.5$ *is accepted, it makes sense to use the conditional prior, given that this event is true, in the inferences. As such it is necessary to condition the prior on the event* $\sum_{i=1}^{m}\left(\sum_{j=1}^{i} p_{Dj}\right) p_{NDi} \leq 1/2$. *In general, it is not clear how to generate from this conditional prior but depending on the size of m and the prior, a brute force approach is to simply generate from the unconditional prior and select those samples for which the condition is satisfied and the same approach works with the posterior.*

Here $m = 5$, *and using uniform priors for* p_{ND} *and* p_D, *the prior probability of* $AUC > 0.5$ *is* 0.281 *while the posterior probability is* 0.998 *so the posterior sampling is much more efficient. Choosing priors that are more favorable to* $AUC > 0.5$ *will improve the efficiency of the prior sampling. Using the conditional priors led to* $AUC(f_{ND}, f_D) = 0.66$ *with* $Pl_{AUC}(f_{ND}, f_D) = [0.60, 0.76]$ *with posterior content* 0.85. *This is similar to the results obtained using the unconditional prior but the conditional prior puts more mass on larger values of the AUC hence the wider plausible region with lower posterior content. Moreover,* $c_{opt}(f_{ND}, f_D) = 2$ *with* $Pl_{c_{opt}}(f_{ND}, f_D) = \{1, 2\}$ *with posterior probability content approximately* 1.00 *(actually* 0.99999*) which reflects virtual certainty that the true optimal value is in* $\{1, 2\}$.

3.3. Binormal Diagnostic

Suppose now that X is a continuous diagnostic variable and it is assumed that the distributions F_D and F_{ND} are normal distributions. The assumption of normality should be checked by an appropriate test and it will be assumed here that this has been carried out and normality was not rejected. While the normality assumption may seem somewhat unrealistic, many aspects of the analysis can be expressed in closed form and this allows for a deeper understanding of ROC analyses more generally.

With Φ denoting the $N(0,1)$ cdf, then $\text{FNR}(c) = \Phi((c-\mu_D)/\sigma_D)$, $\text{FPR}(c) = 1 - \Phi((c-\mu_{ND})/\sigma_{ND})$ so $c = \mu_{ND} + \sigma_{ND}\Phi^{-1}(1-(1-F_{ND}(c)))$ and

$$\text{AUC} = \int_{-\infty}^{\infty} \Phi\left(\frac{\mu_D - \mu_{ND}}{\sigma_D} + \frac{\sigma_{ND}}{\sigma_D} z\right) \varphi(z)\, dz.$$

For given $(\mu_D, \sigma_D, \mu_{ND}, \sigma_{ND})$ and c, all these values can be computed using Φ except the AUC and for that quadrature or simulation via generating $z \sim N(0,1)$ is required.

The following results hold for the AUC with the proofs in the Appendix A.

Lemma 1. *AUC* $> 1/2$ *iff* $\mu_D > \mu_{ND}$ *and when* $\mu_D > \mu_{ND}$, *the AUC is a strictly increasing function of* σ_{ND}/σ_D.

From Lemma 1 it is clear that it makes sense to restrict the parameterization so that $\mu_D > \mu_{ND}$ but we need to test the hypothesis $H_0 : \mu_D > \mu_{ND}$ first. Clearly $\text{Error}(c) = w\text{FNR}(c) + (1-w)\text{FPR}(c) \to 1-w$ as $c \to -\infty$ and $\text{Error}(c) \to w$ as $c \to \infty$ so, if $\text{Error}(c)$ does not achieve a minimum at a finite value of c, then the optimal cut-off is infinite and the optimal error is $\min\{w, 1-w\}$. It is possible to give conditions under which a finite cutoff exists and express c_{opt} in closed form when the parameters and the relevant prevalence w are all known.

Lemma 2. *(i) When* $\sigma_D^2 = \sigma_{ND}^2 = \sigma^2$, *then a finite optimal cut-off minimizing* $\text{Error}(c)$ *exists iff* $\mu_D > \mu_{ND}$ *and in that case*

$$c_{opt} = \frac{\mu_D + \mu_{ND}}{2} + \frac{\sigma^2}{\mu_D - \mu_{ND}} \log\left(\frac{1-w}{w}\right). \tag{5}$$

(ii) When $\sigma_D^2 \neq \sigma_{ND}^2$, *then a finite optimal cut-off exists iff*

$$(\mu_D - \mu_{ND})^2 + 2\left(\sigma_D^2 - \sigma_{ND}^2\right) \log\left(\frac{1-w}{w}\frac{\sigma_D}{\sigma_{ND}}\right) \geq 0 \tag{6}$$

and in that case

$$c_{opt} = \frac{\sigma_{ND}^2\mu_D - \sigma_D^2\mu_{ND}}{\sigma_{ND}^2 - \sigma_D^2} - \frac{\sigma_{ND}\sigma_D}{\sigma_{ND}^2 - \sigma_D^2} \left\{ \begin{array}{c} (\mu_D - \mu_{ND})^2 + \\ 2(\sigma_D^2 - \sigma_{ND}^2) \log\left(\frac{1-w}{w}\frac{\sigma_D}{\sigma_{ND}}\right) \end{array} \right\}^{1/2}. \tag{7}$$

Note that when $w = 1/2$, then in (i) $c_{opt} = (\mu_D + \mu_{ND})/2$ as one might expect. In the case of unequal variances there is an additional restriction beyond $\mu_D \geq \mu_{ND}$ required to hold if the diagnostic is to serve as a reasonable classifier. The following shows that these can be combined in a natural way.

Corollary 1. *The restrictions* $\mu_D \geq \mu_{ND}$ *and (6) hold iff*

$$\mu_D - \mu_{ND} - \left\{\max\left[0, -2\left(\sigma_D^2 - \sigma_{ND}^2\right) \log\left(\frac{1-w}{w}\frac{\sigma_D}{\sigma_{ND}}\right)\right]\right\}^{1/2} \geq 0. \tag{8}$$

So, if one is unwilling to assume constant variance, then the hypothesis $H_0 :$ (8) holds, needs to be assessed. There is some importance to these results as they demonstrate that a finite optimal cutoff may in fact not exist at least when considering both types of error. For example, when $\mu_{ND} = 1, \mu_D = 2, \sigma_D = 1, \sigma_{ND} = 1.5$, then for any $w \leq 0.30885$, the optimal cutoff is $c_{opt} = \infty$ with $\text{Error}(\infty) = w$. When c_{opt} is infinite, then one may need to consider various cutoffs c and find one that is acceptable at least with respect to some of the error characteristics $\text{FNR}(c)$, $\text{FPR}(c)$, $\text{Error}(c)$, $\text{FDR}(c)$ and $\text{FNDR}(c)$.

Consider now examples with equal and unequal variances.

Example 3. *Binormal with* $\sigma_{ND}^2 = \sigma_D^2$.

There may be reasons why the assumption of equal variance is believed to hold but this needs to be assessed and evidence in favor found. If evidence against the assumption is found, then the approach of Example 4 can be used. A possible prior is given by $\pi_1(\mu_{ND}, \sigma^2)\pi_2(\mu_D \mid \sigma^2)$ *where*

$$\mu_{ND} \mid \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2),\ \mu_D \mid \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2),\ 1/\sigma^2 \sim\ gamma(\lambda_1, \lambda_2)$$

which is a conjugate prior. The hyperparameters to be elicited are $(\mu_0, \tau_0^2, \lambda_1, \lambda_2)$. *Consider first eliciting the prior for* (μ_{ND}, σ^2). *For this an interval* (m_1, m_2) *is specified such that is it believed that* $\mu_{ND} \in (m_1, m_2)$ *with virtual certainty (say with probability* $\gamma = 0.99$). *Then putting* $\mu_0 = (m_1 + m_2)/2$ *implies*

$$\gamma \leq \Phi((m_2 - \mu_0)/\tau_0\sigma) - \Phi((m_1 - \mu_0)/\tau_0\sigma) = 2\Phi((m_2 - m_1)/2\tau_0\sigma) - 1$$

which implies $\sigma \leq (m_2 - m_1)/2\tau_0 z_{(1+\gamma)/2}$ *where* $z_{(1+\gamma)/2} = \Phi^{-1}((1+\gamma)/2)$. *The interval* $\mu_{ND} \pm \sigma z_{(1+\gamma)/2}$ *will contain an observation from* F_{ND} *with virtual certainty and let* (l_0, u_0) *be lower and upper bounds on the half-length of this interval so* $l_0/z_{(1+\gamma)/2} \leq \sigma \leq u_0/z_{(1+\gamma)/2}$ *with virtual certainty. This implies* $\tau_0 = (m_2 - m_1)/2u_0$. *This leaves specifying the hyperparameters* (λ_1, λ_2), *and letting* $G(\cdot, \lambda_1, \lambda_2)$ *denote the cdf of the gamma*(λ_1, λ_2) *distribution, then* (λ_1, λ_2) *satisfying*

$$G(z_{(1+\gamma)/2}^2/l_0^2, \lambda_1, \lambda_2) = (1+\gamma)/2,\ G(z_{(1+\gamma)/2}^2/u_0^2, \lambda_1, \lambda_2) = (1-\gamma)/2 \tag{9}$$

will give the specified γ *coverage. Noting that* $G(x, \lambda_1, \lambda_2) = G(\lambda_2 x, \lambda_1, 1)$, *first specify* λ_1 *and solve the first equation in (9) for* λ_2 *and then solve the second equation in (9) for* λ_1 *and continue this iteration until the probability content of* $(l_0/z_{(1+\gamma)/2}, u_0/z_{(1+\gamma)/2})$ *is sufficiently close to* γ. *Using* $s_D^2 = ||x_D - \bar{x}_D 1||^2, s_{ND}^2 = ||x_{ND} - \bar{x}_{ND}1||^2$, *the posterior is then*

$$\begin{aligned}
&\mu_{ND} \mid \sigma^2, x_{ND} \sim N\Big((n_{ND} + 1/\tau_0^2)^{-1}(n_{ND}\bar{x}_{ND} + \mu_0/\tau_0^2), (n_{ND} + 1/\tau_0^2)^{-1}\sigma^2\Big),\\
&\mu_D \mid \sigma^2, x_D \sim N\Big((n_D + 1/\tau_0^2)^{-1}(n_D\bar{x}_D + \mu_0/\tau_0^2), (n_D + 1/\tau_0^2)^{-1}\sigma^2\Big),\\
&1/\sigma^2 \mid (x_{ND}, x_D) \sim\ gamma(\lambda_1 + (n_D + n_{ND})/2, \lambda_x)
\end{aligned}$$

where

$$\begin{aligned}
\lambda_x \ =\ & \lambda_2 + (s_D^2 + s_{ND}^2)/2 + (n_D + 1/\tau_0^2)^{-1}(n_D/\tau_0^2)(\bar{x}_D - \mu_0)^2/2 +\\
& (n_{ND} + 1/\tau_0^2)^{-1}(n_{ND}/\tau_0^2)(\bar{x}_{ND} - \mu_0)^2/2.
\end{aligned}$$

Suppose the following values of the mss were obtained based on samples of $n_{ND} = 25$ *from* $F_{ND} = N(0,1)$ *and* $n_D = 20$ *from* $F_D = N(1,1)$

$$(\bar{x}_{ND}, s_{ND}^2) = (-0.072, 19.638),\ (\bar{x}_D, s_D^2) = (0.976, 16.778).$$

So the true values of the parameters are $\mu_{ND} = 0, \mu_D = 1, \sigma^2 = 1$. *In this case AUC* $= \int_{-\infty}^{\infty} \Phi(1+z)\varphi(z)\,dz = 0.760$. *Supposing that the relevant prevalence is* $w = 0.4, c_{opt} = 0.5 + \log(0.6/0.4) = 0.905$, $FNR(c_{opt}) = \Phi(0.905 - 1) = 0.46$, $FPR(c_{opt}) = 1 - \Phi(0.905) = 0.18$, $Error(c_{opt}) = 0.30, FDR(c_{opt}) = 0.34$, $FNDR(c_{opt}) = 0.27$.

For the prior elicitation, suppose it is known with virtual certainty that both means lie in $(-5, 5)$ *and* $(l_0, u_0) = (1, 10)$ *so we take* $\mu_0 = (-5+5)/2 = 0, \tau_0 = (m_2 - m_1)/2u_0 = 0.5$ *and the iterative process leads to* $(\lambda_1, \lambda_2) = (1.787, 1.056)$. *For inference about* c_{opt} *it is necessary to specify a prior distribution for the prevalence* w. *This can range from* w *being completely known to being completely unknown whence a uniform(0,1) (beta*$(1,1)$*) would be appropriate. Following the developments of Section 3.1, suppose it is known that* $w \in [l, u] = [0.2, 0.6]$ *with prior probability* $\gamma = 0.99$, *so in this case* $\xi_w = (l+u)/2 = 0.4$ *and* $\tau_w = 35.89725$ *and the prior is* $w \sim beta(15.3589, 22.53835)$.

The first inference step is to assess the hypothesis $H_0 : AUC > 1/2$ *which is equivalent to* $H_0 : \mu_{ND} < \mu_D$ *by computing the prior and posterior probabilities of this event to obtain the relative belief ratio. The prior probability of* H_0 *given* σ^2 *is*

$$\int_{-\infty}^{\infty} \Phi((\mu_D - \mu_0)/\tau_0\sigma)(\tau_0\sigma)^{-1}\varphi((\mu_D - \mu_0)/\tau_0\sigma)\, d\mu_D = 1/2$$

and averaging this quantity over the prior for σ^2 *we get* $1/2$*. The posterior probability of this event can be easily obtained via simulating from the joint posterior. When this is done in the specific numerical example, the relative belief ratio of this event is* 2.011 *with posterior content* 0.999 *so there is strong evidence that* $H_0 : AUC > 1/2$ *is true.*

If evidence is found against H_0*, then this would indicate a poor diagnostic. If evidence is found in favor, then we can proceed conditionally given that* H_0 *holds and so condition the joint prior and joint posterior on this event being true when making inferences about AUC,* c_{opt}*, etc. So for the prior it is necessary to generate* $1/\sigma^2 \sim gamma(\alpha_0, \beta_0)$ *and then generate* (μ_D, μ_{ND}) *from the joint conditional prior given* σ^2 *and that* $\mu_D > \mu_{ND}$*. Denoting the conditional priors given* σ^2 *by* $\pi_D(\mu_D \,|\, \sigma^2)$ *and* $\pi_{ND}(\mu_{ND} \,|\, \sigma^2)$*, we see that this joint conditional prior is proportional to*

$$\pi_{ND}(\mu_{ND} \,|\, \sigma^2)\pi_D(\mu_D \,|\, \sigma^2) = \Pi_{ND}(\mu_{ND} < \mu_D \,|\, \mu_D, \sigma^2)\frac{\pi_{ND}(\mu_{ND})}{\Pi_{ND}(\mu_{ND} < \mu_D \,|\, \sigma^2)}\pi_D(\mu_D \,|\, \sigma^2).$$

While generally it is not possible to generate efficiently from this distribution we can use importance sampling to calculate any expectations by generating $\mu_D \sim \mu_D \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2), \mu_{ND} \sim N(\mu_0, \tau_0^2\sigma^2 \,|\, (-\infty, \mu_D])$ *with* $\Pi_{ND}(\mu_{ND} < \mu_D \,|\, \mu_D, \sigma^2) = \Phi((\mu_D - \mu_0)/\tau_0\sigma)$ *serving as the importance sampling weight and where* $N(\mu_0, \tau_0^2\sigma^2 \,|\, (-\infty, \mu_D])$ *denotes the* $N(\mu_0, \tau_0^2\sigma^2)$ *distribution conditioned to* $(-\infty, \mu_D]$ *with density*

$$\Phi^{-1}((\mu_D - \mu_0)/\tau_0\sigma)(2\pi\tau_0^2\sigma^2)^{-1/2}\varphi((\mu_{ND} - \mu_0)/\tau_0\sigma)$$

for $\mu_{ND} \leq \mu_D$ *and 0 otherwise. Generating from this distribution via inversion is easy since the cdf is* $\Phi((\mu_{ND} - \mu_0)/\tau_0\sigma)/\Phi((\mu_D - \mu_0)/\tau_0\sigma)$*. Note that, if we take the posterior from the unconditioned prior and condition that, we will get the same conditioned posterior as when we use the conditioned prior to obtain the posterior. This implies that in the joint posterior for* $(\mu_{ND}, \mu_D, \sigma^2)$ *it is only necessary to adjust the posterior for* μ_{ND} *as was done with the prior and this is also easy to generate from. Note that Lemma 2 (i) implies that it is necessary to use the conditional prior and posterior to guarantee that* c_{opt} *exists finitely.*

Since H_0 *was accepted, the conditional sampling was implemented and the estimate of the AUC is* 0.795 *with plausible region* $[0.670, 0.880]$ *which has posterior content* 0.856*. So the estimate is close to the true value but there is substantial uncertainty. Figure 4 is a plot of the conditioned prior, the conditioned posterior and relative belief ratio for this data.*

With the specified prior for w*, the posterior is beta* $(35.3589, 47.53835)$ *which leads to estimate* 0.444 *for* w *with plausible interval* $(0.374, 0.516)$ *having posterior probability content* 0.782*. Using this prior and posterior for* w *and the conditioned prior and posterior for* $(\mu_D, \mu_{ND}, \sigma^2)$*, we proceed to an inference about* c_{opt} *and the error characteristics associated with this classification. A computational problem arises when obtaining the prior and posterior distributions of* c_{opt} *as it is clear from (5) that these distributions can be extremely long-tailed. As such, we transform to* $c_{\text{mod}} = 0.5 + \arctan(c_{opt})/\pi \in [0, 1]$ *(the Cauchy cdf), obtain the estimate* $c_{\text{mod}}(d)$ *where* $d = (n_{ND}, \bar{x}_{ND}, s_{ND}^2, n_D, \bar{x}_D, s_D^2)$ *and its plausible region and then, applying the inverse transform, obtain* $c_{opt}(d) = \tan(\pi(c_{\text{mod}}(d) - 0.5))$ *and its plausible region. It is notable that relative belief inferences are invariant under 1-1 smooth transformations, so it does not matter which parameterization is used, but it is much easier computationally to work with a bounded quantity. Furthermore, if a shorter tailed cdf is used rather than a Cauchy, e.g., a* $N(0, 1)$ *cdf, then errors can arise due to extreme negative values being always transformed to 0 and very extreme positive values always transformed to 1. Figure 5 is a plot of the prior density, posterior density and relative belief ratio of* c_{mod}*. For these data* $c_{opt}(d) = 0.715$ *with plausible interval* $(0.316, 1.228)$ *having posterior content* 0.860*. Large Monte Carlo samples were used to get smooth estimates of*

the densities and relative belief ratio but these only required a few minutes of computer time on a desktop. The estimated error characteristics at this value of c_{opt} *are as follows:* $FNR(0.715) = 0.41$, $FPR(0.715) = 0.22$, $Error(0.715) = 0.27$, $FDR(0.715) = 0.30$, $FNDR(0.715) = 0.24$ *which are close to the true values.*

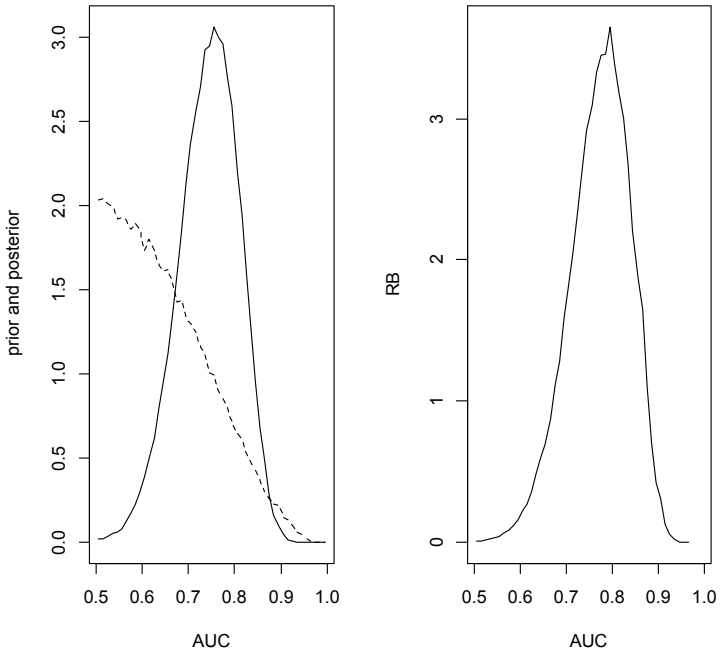

Figure 4. The conditioned prior (- -) and posterior (–) densities (left panel) and the relative belief ratio (right panel) of the AUC in Example 3.

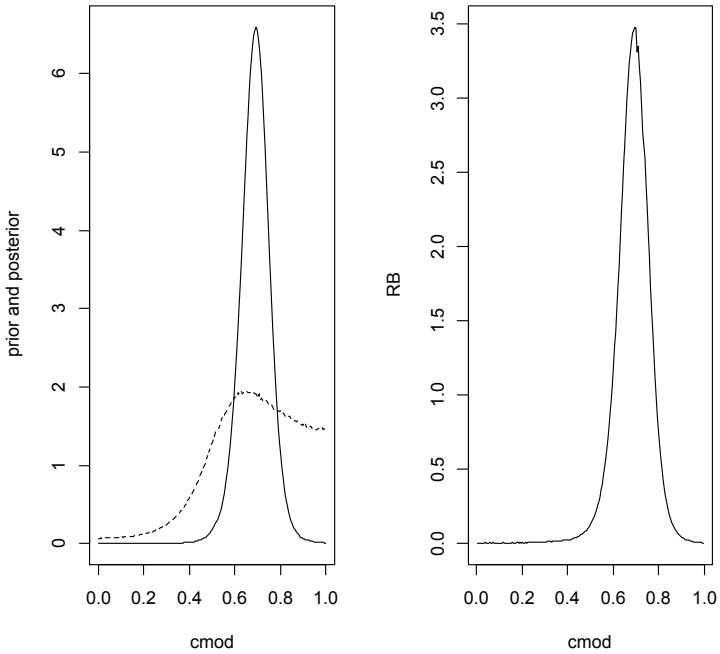

Figure 5. Plots of the prior (- -), posterior (left panel) and relative belief ratio (right panel) of c_{opt} in Example 3.

Example 4. *Binormal with* $\sigma_{ND}^2 \neq \sigma_D^2$.

In this case the prior is given by $\pi_1(\mu_{ND}, \sigma^2_{ND})\pi_2(\mu_D, \sigma^2_D)$ *where*

$$\mu_{ND} \,|\, \sigma^2_{ND} \sim N(\mu_0, \tau_0^2\sigma^2_{ND}),\ 1/\sigma^2_{ND} \sim\ gamma(\lambda_1, \lambda_2)$$
$$\mu_D \,|\, \sigma^2_D \sim N(\mu_0, \tau_0^2\sigma^2_D),\ 1/\sigma^2_D \sim\ gamma(\lambda_1, \lambda_2). \tag{10}$$

Although this specifies the same prior for the two populations, this is easily modified to use different priors and, in any case, the posteriors are different. Again it is necessary to check that the $AUC > 1/2$ *but also to check that* c_{opt} *exists using the full posterior based on this prior and for this we have the hypothesis* H_0 *given by Corollary 1. If evidence in favor of* H_0 *is found, the prior is replaced by the conditional prior given this event for inference about* c_{opt}*. This can be implemented via importance sampling as was done in Example 3 and similarly for the posterior.*

Using the same data and hyperparameters as in Example 3 the relative belief ratio of H_0 *is* 3.748 *with posterior content* 0.828 *so there is reasonably strong evidence in favor of* H_0*. Estimating the value of the AUC is then based on conditioning on* H_0 *being true. Using the conditional prior given that* H_0 *is true, the relative belief estimate of the AUC is* 0.793 *with plausible interval* $(0.683, 0.857)$ *with posterior content* 0.839*. The optimal cutoff is estimated as* $c_{opt}(d) = 0.739$ *with plausible interval* $(0.316, 1.228)$ *having posterior content* 0.875*. Figure 6 is a plot of the prior density, posterior density and relative belief ratio of* c_{mod}*. The estimates of the error characteristics at* $c_{opt}(d)$ *are as follows:* $FNR(0.739) = 0.43$, $FPR(0.739) = 0.19$, $Error(0.739) = 0.28$, $FDR(0.739) = 0.28$, $FNDR(0.624) = 0.264$.

It is notable that these inferences are very similar to those in Example 3. It is also noted that the sample sizes are not big and so the only situation where it might be expected that the inferences will be quite different between the two analyses is when the variances are substantially different.

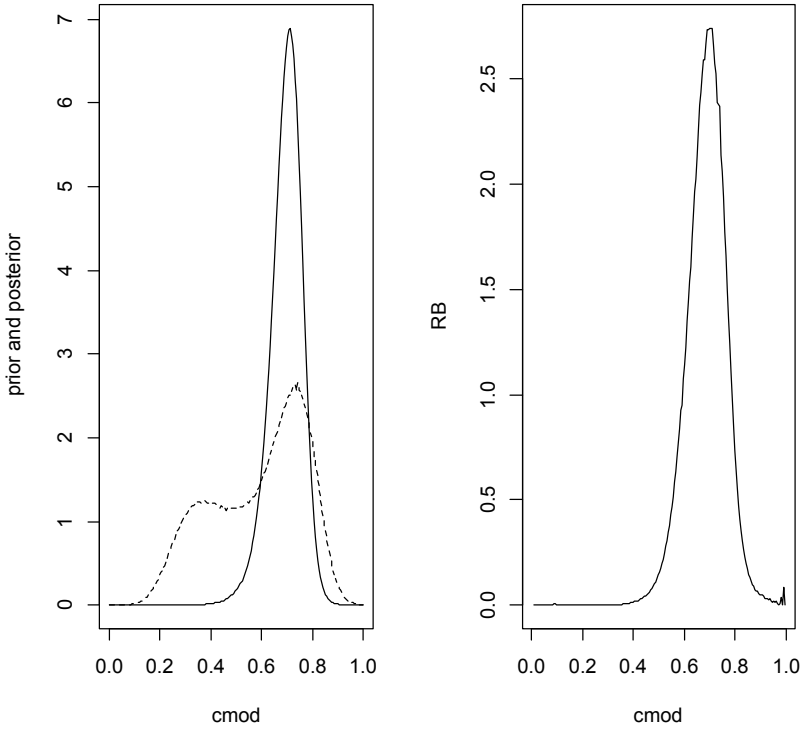

Figure 6. Plots of the prior (- -), posterior (left panel) and relative belief ratio (right panel) of c_{opt} in Example 4.

3.4. Nonparametric Bayes Model

Suppose that X is a continuous variable, of course still measured to some finite accuracy, and available information is such that no particular finite dimensional family of distributions is considered feasible. The situation is considered where a normal distribution $N(\mu, \sigma^2)$, perhaps after transforming the data, is considered as a possible base distribution for X but we want to allow for deviation from this form. Alternative choices can also be

made for the base distribution. The statistical model is then to assume that the x_{ND} and x_D are generated as samples from F_{ND} and F_D, where these are independent values from a DP(a, H) (Dirichlet) process with base $H = N(\mu, \sigma^2)$ for some (μ, σ^2) and concentration parameter a. Actually, since it is difficult to argue for some particular choice of (μ, σ^2), it is supposed that (μ, σ^2) also has a prior $\pi(\mu, \sigma^2)$. The prior on (F_{ND}, F_D) is then specified hierarchically as a mixture Dirichlet process,

$$\begin{aligned}&(\mu_{ND}, \sigma^2_{ND}) \sim \pi \text{ independent of } (\mu_D, \sigma^2_D) \sim \pi,\\ &F_{ND} \,|\, (\mu_{ND}, \sigma^2_{ND}) \sim \text{ DP}(a_{ND}, N(\mu_{ND}, \sigma^2_{ND})) \text{ independent of}\\ &F_D \,|\, (\mu_D, \sigma^2_D) \sim \text{ DP}(a_D, N(\mu_D, \sigma^2_D)).\end{aligned}$$

To complete the prior it is necessary to specify π and the concentration parameters a_{ND} and a_D. For π the prior is taken to be a normal distribution elicited as discussed in Section 3.3 although other choices are possible. For eliciting the concentration parameters, consider how strongly it is believed that normality holds and for convenience suppose $a = a_{ND} = a_D$. If $F \sim$ DP(a, H) with H a probability measure, then $E(F(A)) = H(A)$ and $Var(F(A)) = H(A)(1 - H(A))/(1 + a)$. When F a random measure from P, then $\sup_A P(|F(A) - H(A)| \geq \varepsilon) = \sup_A \{1 - P(\max(0, H(A) - \varepsilon) < F(A) < \min(1, H(A) + \varepsilon))\}$ which, when $P \sim$ DP(a, H), equals

$$\sup_{r \in [0,1]} \{1 - B([\max(0, r - \varepsilon), \min(1, r + \varepsilon)], ar, a(1 - r))\} \tag{11}$$

where $B(\cdot, \beta_1, \beta_2)$ denotes the beta(β_1, β_2) measure. This upper bound on the probability that the random F differs from H by at least ε on an event can be made as small as desirable by choosing a large enough. For example, if $\varepsilon = 0.25$ and it is required that this upper bound be less than 0.1, then this satisfied when $a \geq 9.8$ and if instead $\varepsilon = 0.1$, then $a \geq 66.8$ is necessary. Note that, since this bound holds for every continuous probability measure H, it also holds when H is random, as considered here. So a is controlling how close it is believed that the true distribution is to H. Alternative methods for eliciting a can be found in [24,25].

Generating (F_{ND}, F_D) from the prior for given (a, H) can only be done approximately and the approach of [26] is adopted. For this, integer n^* is specified and measure $P_{n^*} = \sum_{i=1}^{n^*} p_{i,n^*} I_{\{c_i\}}$ is generated where $(p_{1,n^*}, \ldots, p_{n^*,n^*}) \sim Dirichlet(a/n^*, \ldots, .a/n^*)$ independent of $c_1, \ldots, c_{n^*} \overset{iid}{\sim} H$, since $P_{n^*} \overset{w}{\to}$ DP(a, H) as $n^* \to \infty$. So to carry out a priori calculations proceed as follows. Generate

$$\begin{aligned}&(p_{ND1,n^*}, \ldots, p_{NDn^*,n^*}) \sim \text{Dirichlet}((a/n^*)\mathbf{1}_{n^*}),\ (\mu_{ND}, \sigma^2_{ND}) \sim \pi,\\ &(c_{ND1}, \ldots, c_{NDn^*}) \,|\, (\mu_{ND}, \sigma^2_{ND}) \overset{i.i.d.}{\sim} N(\mu_{ND}, \sigma^2_{ND}),\ w \sim \text{beta}(\alpha_{1w}, \alpha_{2w})\end{aligned}$$

and similarly for $(p_{D1,n^*}, \ldots, p_{Dn^*,n^*})$, (μ_D, σ^2_D), and $(c_{D1}, \ldots, c_{Dn^*})$. Then $F_{ND,n^*}(c) = \sum_{\{i: c_{NDi} \leq c\}} p_{NDin^*}$ is the random cdf at $c \in R^1$ and similarly for F_{D,n^*}, so AUC $= \sum_{i=1}^{n^*} (1 - F_{D,n^*}(c_{NDi})) p_{NDi,n^*}$ is a value from the prior distribution of the AUC. This is done repeatedly to get the prior distribution of the AUC as in our previous discussions and we proceed similarly for the other quantities of interest.

Now $F_{ND} \,|\, x_{ND}, (\mu_{ND}, \sigma^2_{ND}, \mu_D, \sigma^2_D) \sim$ DP$(a + n_{ND}, H_{ND})$ independent of $F_D \,|\, x_D$, $(\mu_{ND}, \sigma^2_{ND}, \mu_D, \sigma^2_D) \sim$ DP$(a + n_D, H_D)$ with $H_{ND}(c) = a\Phi((c - \mu_{ND})/\sigma_{ND})/(a + n_{ND}) + n_{ND}\hat{F}_{ND}(c)/(a + n_{ND})$ and $\hat{F}_{ND}(c) = \sum_{i=1}^{n_{ND}} I_{(-\infty, c]}(x_{NDi})/n_{ND}$ is the empirical cdf (ecdf) based on x_{ND} and similarly for H_D. The posteriors of $(\mu_{ND}, \sigma^2_{ND})$ and (μ_D, σ^2_D) are obtained via results in [27,28]. The posterior density of $(\mu_{ND}, \sigma^2_{ND})$ given x_{ND} is proportional to

$$\pi(\mu_{ND}, \sigma^2_{ND}) \prod_{i=1}^{\tilde{n}_{ND}} \sigma^{-1}_{ND} \varphi((\tilde{x}_{NDi} - \mu_{ND})/\mu_{ND})$$

where $\tilde{n}_{ND}$ is the number of unique values in x_{ND} and $\{\tilde{x}_{ND1}, \ldots, \tilde{x}_{ND\tilde{n}_{ND}}\}$ is the set of unique values with mean $\tilde{x}_{ND}$ and sum of squared deviations $\tilde{s}^2_{ND}$. From this it is immediate that

$$\mu_{ND} \,|\, \sigma^2_{ND}, x_{ND} \sim N\Big((\tilde{n}_{ND} + 1/\tau_0^2)^{-1}(\tilde{n}_{ND}\tilde{x}_{ND} + \mu_0/\tau_0^2), (\tilde{n}_{NDND} + 1/\tau_0^2)^{-1}\sigma^2_{ND}\Big),$$
$$1/\sigma^2_{ND} \,|\, x_{ND} \sim \text{ gamma}(\alpha_0 + \tilde{n}_{ND}/2, \tilde{\lambda}_{x_{ND}})$$

where $\tilde{\lambda}_{x_{ND}} = \lambda_0 + \tilde{s}^2_{ND}/2 + (\tilde{n}_{ND} + 1/\tau_0^2)^{-1}(\tilde{n}_{ND}/\tau_0^2)(\tilde{x}_{ND} - \mu_0)^2/2$. A similar result holds for the posterior of (μ_D, σ^2_D).

To approximately generate from the full posterior specify some n^{**}, put $p_{a,n_{ND}} = a/(a + n_{ND}), q_{a,n_{ND}} = 1 - p_{a,n_{ND}}$ and generate

$$(p_{ND1,n^{**}}, \ldots, p_{NDn^{**},n^{**}}) \,|\, x_{ND} \sim \text{ Dirichlet}(((a + n_{ND})/n^{**})\mathbf{1}_{n^{**}}),$$
$$(\mu_{ND}, \sigma^2_{ND}) \,|\, x_{ND} \sim \pi(\cdot \,|\, x_{ND}),$$
$$(c_{ND1}, \ldots, c_{NDn^{**}}) \,|\, (\mu_{ND}, \sigma^2_{ND}), x_{ND} \overset{i.i.d.}{\sim} p_{a,n_{ND}} N(\mu_{ND}, \sigma^2_{ND}) + q_{a,n_{ND}} \hat{F}_{ND},$$
$$w \,|\, x_{ND} \sim \text{ beta}(\alpha_{1w} + n_D, \alpha_{2w} + n_{ND})$$

and similarly for $(p_{D1,n^{**}}, \ldots, p_{Dn^{**},n^{**}})$, (μ_D, σ^2_D) and $(c_{D1}, \ldots, c_{Dn^{**}})$. If the data does not comprise a sample from the full population, then the posterior for w is replaced by its prior.

There is an issue that arises when making inference about c_{opt}, namely, the distributions for c_{opt} that arises from this approach can be very irregular and particularly the posterior distribution. In part this is due to the discreteness of the posterior distributions of F_{ND} and F_D . This does not affect the prior distribution because the points on which the generated distributions are concentrated vary quite continuously among the realizations and this leads to a relatively smooth prior density for c_{opt}. For the posterior, however, the sampling from the ecdf leads to a very irregular, multimodal density for c_{opt}. So some smoothing is necessary in this case.

Consider now applying such an analysis to the dataset of Example 3, where we know the true values of the quantities of interest and then to a dataset concerned with the COVID-19 epidemic.

Example 5. *Binormal data (Examples 3 and 4)*

The data used in Example 3 are now analyzed but using the methods of this section. The prior on $(\mu_{ND}, \sigma^2_{ND}), (\mu_D, \sigma^2_D)$ *and* w *is taken to be the same as that used in Example 4 so the variances are not assumed to be the same. The value* $\varepsilon = 0.25$ *is used and requiring (11) to be less than* 0.018 *leads to* $a = 20$. *So the true distributions are allowed to differ quite substantially from a normal distribution. Testing the hypothesis* $H_0 : AUC > 1/2$ *led to the relative belief ratio* 1.992 *(maximum possible value is* 2*) and the strength of the evidence is* 0.997 *so there is strong evidence that* H_0 *is true. The AUC, based on the prior conditioned on* H_0 *being true, is estimated to be equal to* 0.839 *with plausible interval* $(0.691, 0.929)$ *having posterior content* 0.814. *For these data* $c_{opt}(d) = 0.850$ *with plausible interval* $(0.45, 1.75)$ *having posterior content* 0.835. *The true value of the AUC is* 0.760 *and the true value of* c_{opt} *is* 0.905 *so these inferences are certainly reasonable although, as one might expect, when the length of the plausible intervals are taken into account, they are not as accurate as those when binormality is assumed as this is correct for this data. So the DP approach worked here although the posterior density for* c_{opt} *was quite multimodal and required some smoothing (averaging 3 consecutive values).*

Example 6. *COVID-19 data.*

A dataset was downloaded from https://github.com/YasinKhc/Covid-19 containing data on 3397 individuals diagnosed with COVID-19 and includes whether or not the patient survived the disease, their gender and their age. There are 1136 complete cases on these variables of which 646 are male, with 52 having died, and 490 are female, with 25 having died. Our interest is in the use of a patient's age X to predict whether or not they will survive. More detail on this dataset

can be found in [29]. The goal is to determine a cutoff age so that extra medical attention can be paid to patients beyond that age. Furthermore, it is desirable to see whether or not gender leads to differences so separate analyses can be carried out by gender. So, for example, in the male group ND refers to those males with COVID-19 that will not die and D refers to the population that will. Looking at histograms of the data, it is quite clear that binormality is not a suitable assumption and no transformation of the age variable seems to be available to make a normality assumption more suitable. Table 3 gives summary statistics for the subgroups. Of some note is that condition (8), when using standard estimates for population quantities such as $w = 52/646 = 0.08$ *for Males and* $w = 25/490 = 0.05$ *for females, is not satisfied which suggests that in a binormal analysis no finite optimal cutoff exists.*

Table 3. Summary statistics for the data in Example 6.

Group	Number	Mean	std. dev.	Min	Max
ND males	594	48.81	17.72	0.50	85.00
D males	52	68.46	13.66	36.00	89.00
ND females	465	48.69	18.73	2.00	96.00
D females	25	77.36	12.12	48.00	95.00

For the prior, it is assumed that $(\mu_{ND}, \sigma^2_{ND})$ *and* (μ_D, σ^2_D) *are independent values from the same prior distribution as in (10). For the prior elicitation, as discussed in Example 3, suppose it is known with virtual certainty that both means lie in* $(20, 70)$ *and* $(l_0, u_0) = (20, 50)$ *so we take* $\mu_0 = 45$, $\tau_0 = (m_2 - m_1)/2u_0 = 0.75$ *and the iterative process leads to* $(\lambda_1, \lambda_2) = (8.545, 1080.596)$ *which implies a prior on the* σ*'s with mode at* 10.932 *and the interval* $(7.764, 19.411)$ *containing* 0.99 *of the prior probability. Here the relevant prevalence refers to the proportion of COVID-19 patients that will die and it is supposed that* $w \in [0.00, 0.15]$ *with virtual certainty which implies* $w \sim beta(9.81, 109.66)$*. So the prior probability that someone with COVID-19 will die is assumed to be less than 15% with virtual certainty. Since normality is not an appropriate assumption for the distribution of X, the choice* $\varepsilon = 0.25$ *with the upper bound (11) equal to* 0.1 *seems reasonable and so* $a = 9.8$*. This specifies the prior that is used for the analysis with both genders and it is to be noted that it is not highly informative.*

For males the hypothesis $AUC > 1/2$ *is assessed and* $RB = 1.991$ *(maximum value 2) with strength effectively equal to* 1.00 *was obtained, so there is extremely strong evidence that this is true. The unconditional estimate of the AUC is* 0.808 *with plausible region* $[0.698, 0.888]$ *having posterior content* 0.959*, so there is a fair bit of uncertainty concerning the true value. For the conditional analysis, given that* $AUC > 1/2$*, the estimate of the AUC is* 0.806 *with plausible region* $[0.731, 0.861]$ *having posterior content* 0.932*. So the conditional analysis gives a similar estimate for the AUC with a small increase in accuracy. In either case it seems that the AUC is indicating that age should be a reasonable diagnostic. Note that the standard nonparametric estimate of the AUC is* 0.810 *so the two approaches agree here. For females the hypothesis* $AUC > 1/2$ *is assessed and* $RB = 1.994$ *with strength effectively equal to* 1 *was obtained, so there is extremely strong evidence that this is true. The unconditional estimate of the AUC is* 0.873 *with plausible region* $(0.742, 0.948)$ *having posterior content* 0.968*. For the conditional analysis, given that* $AUC > 1/2$*, the estimate of the AUC is* 0.874 *with plausible region* $(0.791, 0.936)$ *having posterior content* 0.956*. The traditional estimate of the AUC is* 0.902 *so the two approaches are again in close agreement.*

Inferences for c_{opt} *are more problematical in both genders. Consider the male data. The data set is very discrete as there are many repeats and the approach samples from the ecdf about 84% of the time for the males that died and 98% of the time for the males that did not die. The result is a plausible region that is not contiguous even with smoothing. Without smoothing the estimate is* $c_{opt}(d) = 85.5$ *for males, which is a very dominant peak for the relative belief ratio. The plausible region contains* 0.928 *of the posterior probability and, although it is not a contiguous interval, the subinterval* $[85.2, 85.8]$ *is a* 0.58*-credible interval for* c_{opt} *that is in agreement with the evidence. If we make the data continuous by adding a uniform(0,1) random error to each age in the data set, then* $c_{opt}(d) = 86.1$ *and plausible interval* $[75.9, 86.7]$ *with posterior content* 0.968 *is obtained. These cutoffs are both greater than the maximum value in the ND data, so there is ample protection*

against false positives but it is undoubtedly false negatives that are of most concern in this context. If instead the FNDR is used as the error criterion to minimize, then $c_{opt}(d) = 35.7$ *and plausible interval* $[26.1, 35.7]$ *with posterior content* 0.826 *is obtained and so in this case there will be too many false positives. So a useful optimal cutoff incorporating the relevant prevalence does not seem to exist with these data.*

If the relevant prevalence is ignored and $w_0 FNR + (1 - w_0) FPR$ *is used for some fixed weight* w_0 *to determine* $c_{opt}(d)$*, then more reasonable values are obtained. Table 4 gives the estimates for various* w_0 *values. With* $w_0 = 0.5$ *(corresponding to using Youden's index)* $c_{opt}(d) = 65.7$ *while if* $w_0 = 0.7$*, then* $c_{opt}(d) = 56.7$*. When* w_0 *is too small or too large then the value of* $c_{opt}(d)$ *is not useful. While these estimates do not depend on the relevant prevalence, the error characteristics that do depend on this prevalence (as expressed via its prior and posterior distributions) can still be quoted and a decision made as to whether or not to use the diagnostic. Table 5 contains the estimates of the error characteristics at* $c_{opt}(d)$ *for various values of* w_0 *where these are determined using the prior and posterior on the relevant prevalence* w*. Note that these estimates are determined as the values that maximize the corresponding relative belief ratios and take into account the posterior of* w*. So, for example, the estimate of the Error is not the convex combination of the estimates of FNR and FPR based on the* w_0 *weight. Another approach is to simply set the cutoff Age at a value at a value* c_0 *and then investigate the error characteristics at that value. For example, with* $c_0 = 60$*, then the estimated values are given by* $FNR(c_0) = 0.238$*,* $FPR(c_0) = 0.308$*,* $Error(c_0) = 0.328$*,* $FDR(c_0) = 0.818$ *and* $FNDR(c_0) = 0.028$*.*

Similar results are obtained for the cutoff with female data although with different values. Overall, Age by itself does not seem to be useful classifier although that is a decision for medical practitioners. Perhaps it is more important to treat those who stand a significant chance of dying more extensively and not worry too much that some treatments are not necessary. The clear message from this data, however, is that a relatively high AUC does not immediately imply that a diagnostic is useful and the relevant prevalence is a key aspect of this determination.

Table 4. Weighted error w_0FNR+$(1 - w_0)$FPR determining $c_{opt}(d)$ for Males in Example 6.

w_0 = Weight of FNR	$c_{opt}(d)$	Plausible Range (post. prob.)
0.1	85.5	75.3–118.5 (0.945)
0.3	65.1	64.5–85.5 (0.868)
0.5	65.1	55.5–72.3 (0.939)
0.7	56.7	35.7–58.5 (0.919)
0.9	35.7	33.3–52.5 (0.875)

Table 5. Error characteristics for Males in Example 6 at various weights.

w_0 = Weight of FNR	FNR	FPR	Error	FDR	FNDR
0.1	0.918	0.008	0.008	0.458	0.073
0.3	0.368	0.183	0.213	0.733	0.043
0.5	0.368	0.183	0.213	0.733	0.038
0.7	0.158	0.358	0.363	0.823	0.018
0.9	0.003	0.753	0.688	0.893	0.003

4. Conclusions

ROC analyses represent a significant practical application of statistical methodology. While previous work has considered such analyses within a Bayesian framework, this has typically required the specification of loss functions. Losses are not required in the approach taken here which is entirely based on a natural characterization of statistical evidence via the principle of evidence and the relative belief ratio. As discussed in Section 2.2 this results in a number of good properties for the inferences that are not possessed by inferences derived by other approaches. While the Bayes factor is also a valid measure of evidence, its usage is far more restricted than the relative belief ratio which can be applied with any prior, without the need for any modifications, for both hypothesis assessment and

estimation problems. This paper has demonstrated the application of relative belief to ROC analyses under a number of model assumptions. In addition, as documented in points (ii)–(vi) of the Introduction, a number of new results have been developed for ROC analyses more generally.

Author Contributions: Methodology, L.A.-L. and M.E.; Investigation, Q.L.; Writing—original draft, M.E.; Supervision, M.E. All authors have read and agreed to the published version of the manuscript.

Funding: Evans was supported by grant 10671 from the Natural Sciences and Engineering Research Council of Canada.

Data Availability Statement: The data and R code used for the examples in Sections 3.2–3.4 can be obtained at https://utstat.utoronto.ca/mikevans/software/ROCcodeforexamples.zip (accessed on 15 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Lemma 1. Consider $\int_{-\infty}^{\infty}\Phi(a+bz)\varphi(z)\,dz$ as a function of b, so

$$\begin{aligned}&\frac{d}{db}\int_{-\infty}^{\infty}\Phi(a+bz)\varphi(z)\,dz=\int_{-\infty}^{\infty}z\varphi(a+bz)\varphi(z)\,dz\\&=\frac{1}{\sqrt{2\pi}\sqrt{1+b^2}}\exp\left(-\frac{a^2}{2(1+b^2)}\right)\int_{-\infty}^{\infty}z\sqrt{1+b^2}\varphi(\sqrt{1+b^2}(z-(1+b^2)^{-1}ab))\,dz\\&=\frac{1}{\sqrt{2\pi}\sqrt{1+b^2}}\exp\left(-\frac{a^2}{2(1+b^2)}\right)\frac{ab}{1+b^2}.\end{aligned}$$

When $a>0$, then $\int_{-\infty}^{\infty}\Phi(a+bz)\varphi(z)\,dz$ is increasing in b for $b>0$, decreasing in b for $b<0$, equals 0 when $b=0$ and when $a<0$ it is decreasing in b for $b>0$, increasing in b for $b<0$. Therefore, when $a>0, b>0$, then $\int_{-\infty}^{\infty}\Phi(a+bz)\varphi(z)\,dz\geq\Phi(a)>1/2$ and when $a\leq 0, b>0$ then $\int_{-\infty}^{\infty}\Phi(a+bz)\varphi(z)\,dz\leq\Phi(a)\leq 1/2$. □

Proof of Lemma 2. Note that c_{opt} will satisfy

$$\frac{d}{dc}\text{Error}(c)=\frac{w}{\sigma_D}\varphi\left(\frac{c-\mu_D}{\sigma_D}\right)-\frac{1-w}{\sigma_{ND}}\varphi\left(\frac{c-\mu_{ND}}{\sigma_{ND}}\right)=0$$

which implies

$$\varphi\left(\frac{c-\mu_D}{\sigma_D}\right)/\varphi\left(\frac{c-\mu_{ND}}{\sigma_{ND}}\right)=\frac{1-w}{w}\frac{\sigma_D}{\sigma_{ND}}\tag{A1}$$

So c_{opt} is a root of the quadratic $(1/\sigma_D^2-1/\sigma_{ND}^2)c^2-2(\mu_D/\sigma_D^2-\mu_{ND}/\sigma_{ND}^2)c+(\mu_D^2/\sigma_D^2-\mu_{ND}^2/\sigma_{ND}^2+2\log((1-w)\sigma_D/w\sigma_{ND}))$. A single real root exists when $\sigma_D^2=\sigma_{ND}^2=\sigma^2$ and is given by (5).

If $\sigma_D^2\neq\sigma_{ND}^2$, then there are two real roots when the discriminant

$$\begin{aligned}&4(\mu_D/\sigma_D^2-\mu_{ND}/\sigma_{ND}^2)^2-4(1/\sigma_D^2-1/\sigma_{ND}^2)(\mu_D^2/\sigma_D^2-\mu_{ND}^2/\sigma_{ND}^2+\\&2\log((1-w)\sigma_D/w\sigma_{ND}))\geq 0\end{aligned}$$

establishing (6). To be a minimum the root c has to satisfy

$$0<\frac{d^2\text{Error}_w(c)}{dc^2}=-\frac{w}{\sigma_D^2}\left(\frac{c-\mu_D}{\sigma_D}\right)\varphi\left(\frac{c-\mu_D}{\sigma_D}\right)+\frac{1-w}{\sigma_{ND}^2}\left(\frac{c-\mu_{ND}}{\sigma_{ND}}\right)\varphi\left(\frac{c-\mu_{ND}}{\sigma_{ND}}\right)$$

and by (A1), this holds iff

$$0 < -\frac{w}{\sigma_D^2}\left(\frac{c-\mu_D}{\sigma_D}\right)\frac{1-w}{w}\frac{\sigma_D}{\sigma_{ND}} + \frac{1-w}{\sigma_{ND}^2}\left(\frac{c-\mu_{ND}}{\sigma_{ND}}\right) = \frac{1-w}{\sigma_{ND}}\left\{\frac{c-\mu_{ND}}{\sigma_{ND}^2} - \frac{c-\mu_D}{\sigma_D^2}\right\}$$

which is true iff $(1/\sigma_D^2 - 1/\sigma_{ND}^2)c < \mu_D/\sigma_D^2 - \mu_{ND}/\sigma_{ND}^2$. When $\sigma_D^2 = \sigma_{ND}^2$ this is true iff $\mu_D > \mu_{ND}$ which completes the proof of (i). When $\sigma_D^2 \neq \sigma_{ND}^2$ this, together with the formula for the roots of a quadratic establishes (7). □

Proof of Corollary 1. Suppose $\mu_D \geq \mu_{ND}$ and (6) hold. Then putting

$$a = 2\left(\sigma_D^2 - \sigma_{ND}^2\right)\log((1-w)w^{-1}\sigma_D\sigma_{ND}^{-1})$$

we have that, for fixed $\mu_D, \sigma_D^2, \sigma_{ND}^2$ and w, then $(\mu_D - \mu_{ND})^2 + a$ is a quadratic in μ_{ND}. This quadratic has discriminant $-4a$ and so has no real roots whenever $a > 0$ and, noting a does not depend on μ_D, the only restriction on μ_{ND} is $\mu_{ND} \leq \mu_D$. When $a \leq 0$ the roots of the quadratic are given by $\mu_D \pm \sqrt{-a}$ and so, since the quadratic is negative between the roots and $\mu_D - \sqrt{-a} \leq \mu_D \leq \mu_D + \sqrt{-a}$ the two restrictions imply $\mu_{ND} \leq \mu_D - \sqrt{-a}$. Combining the two cases gives (8).

Now suppose (8) holds. Then $\mu_{ND} \leq \mu_D - \{\max(0,-a)\}^{1/2} \leq \mu_D$ which gives the first restriction and also $\mu_{ND} - \mu_D \leq -\{\max(0,-a)\}^{1/2} \leq 0$ which implies $(\mu_{ND} - \mu_D)^2 \geq \max(0,-a)$ and so $(\mu_{ND} - \mu_D)^2 + a \geq \max(0,-a) + a$ and by examining the cases $a \leq 0$ and $a > 0$ we conclude that (6) holds. □

References

1. Metz, C.; Pan, X. "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *Math. Psychol.* **1999**, *43*, 1–33. [CrossRef] [PubMed]
2. Perkins, N.J.; Schisterman, E.F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the Receiver Operating Characteristic Curve. *Am. J. Epidemiol.* **2006**, *163*, 670–675. [CrossRef] [PubMed]
3. López-Ratón, M.; Rodríguez-Álvarez, M.X.; Cadarso-Suárez, C.; Gude-Sampedro, F. OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *J. Stat.* **2014**, *61*, 8. [CrossRef]
4. Unal, I. Defining an optimal cut-point value in ROC analysis: An alternative approach. *Comput. Math. Model. Med.* **2017**, *2017*, 3762651. [CrossRef]
5. Verbakel, J.Y.; Steyerberg, E.W.; Uno, H.; De Cock, B.; Wynants, L.; Collins, G.S.; Van Calster, B. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Clin. Epidemiol.* **2020**. *in press*.
6. Hand, D. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Mach. Learn.* **2009**, *99*, 103–123. [CrossRef]
7. Evans, M. Measuring Statistical Evidence Using Relative Belief. In *Monographs on Statistics and Applied Probability*; CRC Press: Boca Raton, FL, USA; Taylor & Francis: Abingdon, UK, 2015; Volume 144.
8. O'Malley, A.J.; Zou, K.H.; Fielding, J.R.; Tempany, C.M.C. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: Prostate biopsy and spiral CT of ureteral stiones. *Acad. Radiol.* **2001**, *8*, 5407–5420.
9. Gu, J.; Ghosal, S.; Roy, A. Bayesian bootstrap estimation of ROC curve. *Stat. Med.* **2008**, *27*, 5407–5420. [CrossRef]
10. Erkanli, A.; Sung, M.; Costello, E.J.; Angold, A. Bayesian semi-parametric ROC analysis. *Stat. Med.* **2006**, *25*, 3905–3928. [CrossRef] [PubMed]
11. de Carvalho, V.; Jara, A.; Hanson, E.; de Carvalho, M. Bayesian nonparametric ROC regression modeling. *Bayesian Anal.* **2013**, *3*, 623–646. [CrossRef]
12. Ladouceur, M.; Rahme, E.; Belisle, P.; Scott, A.; Schwartzman, K.; Joseph, L. Modeling continuous diagnostic test data using approximate Dirichlet process distributions. *Stat. Med.* **2011**, *30*, 2648–2662. [CrossRef] [PubMed]
13. Christensen, R.; Johnson, W.; Branscum, A.; Hanson, T.E. *Bayesian Ideas and Data Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2011.
14. Rosner, G.L.; Laud, P.W.; Johnson, W.O. *Bayesian Thinking in Biostatistics*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2021.
15. Diab, A.; Hassan, M.; Marquea, C.; Karlsson, B. Performance analysis of four nonlinearity analysis methods using a model with variable complexity and application to uterine EMG signals. *Med. Eng. Phys.* **2014**, *36*, 761–767. [CrossRef] [PubMed]
16. Gao, X.-Y.; Guo, Y.-J.; Shan, W.-R. Regarding the shallow water in an ocean via a Whitham-Broer-Kaup-like system: Hetero-Bäcklund transformations, bilinear forms and M solitons. *Chaos Solitons Fractals* **2022**, *162*, 112486 [CrossRef]

17. Gao, X.-T.; Tian, B. Water-wave studies on a (2+1)-dimensional generalized variable-coefficient Boiti–Leon–Pempinelli system. *Appl. Math. Lett.* **2022**, *128*, 107858. [CrossRef]
18. Obuchowski, N.; Bullen, J. Receiver operating characteristic (ROC) curves: Review of methods with applications in diagnostic medicine. *Phys. Med. Biol.* **2018**, *63*, 07TR01. [CrossRef]
19. Zhou, X.; Obuchowski, N.; McClish, D. *Statistical Methods in Diagnostic Medicine*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2011.
20. Al-Labadi, L.; Evans, M. Optimal robustness results for some Bayesian procedures and the relationship to prior-data conflict. *Bayesian Anal.* **2017**, *12*, 702–728. [CrossRef]
21. Gu, Y.; Li, W.; Evans, M.; Englert, B.-G. Very strong evidence in favor of quantum mechanics and against local hidden variables from a Bayesian analysis. *Phys. Rev. A* **2019**, *99*, 022112. [CrossRef]
22. Englert, B.-G.; Evans, M.; Jang, G.-H.; Ng, H.-K.; Nott, D.; Seah, Y.-L. Checking the model and the prior for the constrained multinomial. *arXiv* **2018**, arXiv:1804.06906.
23. Evans, M.; Guttman, I.; Li, P. Prior elicitation, assessment and inference with a Dirichlet prior. *Entropy* **2017**, *19*, 564. [CrossRef]
24. Swartz, T. Subjective priors for the Dirichlet process. *Commun. Stat. Theory Methods* **1993**, *28*, 2821–2841. [CrossRef]
25. Swartz, T. Nonparametric goodness-of-fit. *Commun. Stat. Theory Methods* **1999**, *22*, 2999–3011. [CrossRef]
26. Ishwaran, H.; Zarepour, M. Exact and approximate sum representations for the Dirichlet process. *Can. J. Stat.* **2002**, *30*, 269–283. [CrossRef]
27. Antoniak, C.E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* **1974**, *2*, 1152–1174. [CrossRef]
28. Doss, H. Bayesian Nonparametric Estimation for Incomplete Data Via Successive Substitution Sampling. *Ann. Stat.* **1994**, *22*, 1763–1786. [CrossRef]
29. Charvadeh, Y.K.; Yi, G.Y. Data visualization and descriptive analysis for understanding epidemiological characteristics of COVID-19: A case study of a dataset from January 22, 2020 to March 29, 2020. *J. Data Sci.* **2020**, *18*, 526–535. [CrossRef]