STATISTICS IN MEDICINE
Statist. Med. 2014; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Bayesian ROC curve estimation under verification bias

Jiezhun Gu[1,*] Subhashis Ghosal[2] and David E. Kleiner[3]

[1] *Duke Clinical Research Institute, Duke University Medical Center, P.O. Box 17969, Durham, NC, 27715*
[2] *Department of Statistics, North Carolina State University, Raleigh, NC 27695*
[3] *Laboratory of Pathology,National Cancer Institute, Bethesda, MD, 20892*

SUMMARY

Receiver Operating Characteristic (ROC) curve has been widely used in medical science for its ability to measure the accuracy of diagnostic tests under the gold standard. However, in a complicated medical practice, a gold standard test can be invasive, expensive, and its result may not always be available for all the subjects under study. Thus a gold standard test is implemented only when it is necessary and possible. This leads to the so called "verification bias", meaning that subjects with verified disease status (also called label) are not selected in a completely random fashion. In this paper, we propose a new Bayesian approach for estimating an ROC curve based on continuous data following the popular semiparametric binormal model in the presence of verification bias. By using a rank-based likelihood, and following Gibbs sampling techniques, we compute the posterior distribution of the binormal parameters intercept and slope, as well as the area under the curve (AUC) by imputing the missing labels within Markov Chain Monte-Carlo (MCMC) iterations. Consistency of the resulting posterior under mild conditions is also established. We compare the new method with other comparable methods and conclude that our estimator performs well in terms of accuracy.

Keywords: Binormal model; MAR assumption; Posterior consistency; ROC curve; Verification bias-correction.

1. Introduction

The Receiver Operating Characteristic (ROC) curve for a long time has been widely used in diagnostic medicine [1] because of its ability to incorporate accuracy of all decision rules in a curve plotted in the unit square. When the true disease status of each study subject is known by the most accurate diagnostic test called the gold standard test, the ROC curve has been used to compare the accuracy of other available diagnostic test(s) to the gold standard test. The ROC curve is the plot of the true positive rate (abbreviated as TPR, also called sensitivity) versus the false positive rate (abbreviated as FPR, also called one minus specificity) by varying a decision threshold value c. The decision threshold value c is used to determine the diagnostic result as positive, or negative, depending on whether the test is not less than, or less than c, respectively.

*Correspondence to: Duke Clinical Research Institute, Duke University Medical Center, P.O. Box 17969, Durham, NC, 27715. Email: jiezhun.gu@duke.edu.

Received
Revised

Since gold standard tests may be invasive and expensive, it is more ethical that verification of the true disease status of study subjects would generally be obtained only for high risk subjects according to the screening test. For example, liver biopsy is considered as the gold standard in evaluating chronic hepatitis and fibrosis. It costed on an average over US $1000 in 2004 without complications and about US $3000 with complications [2]. Further, liver biopsy procedure is invasive. Complications of liver biopsy include significant bleeding and hospitalization. Fatal complications have been reported up to 0.038% among the biopsy patients [3].

Because of a differential in the chance of verification between subjects with high and low risk, it follows that the verified subjects are not sampled randomly from the population. Hence, an estimator of the accuracy of a diagnostic test given by the area under the ROC curve (AUC) based on only the subjects with labels may be biased. This is known as the verification bias. Correcting for the verification bias involves dealing with the missing data. Here, we use the commonly used assumption for missing verification of disease status, assuming missing at random (MAR) introduced by Little and Rubin [4], which means the chance of missing the verification of disease status is independent of the disease itself conditional on the observed measurements.

In this paper, we focus on the problem of estimating diagnostic accuracy of a single test in the presence of verification bias. When the screening test is binary or ordinal, under the MAR assumption, Begg and Greenes [5] proposed an adjusted estimate of TPR and FPR. Other more efficient methods are discussed by Reilly and Pepe [6], Zhou [7], Clayton et al. [8]. Gastwirth, Johnson and Reneau [9] adopted a Bayesian analysis to estimate sensitivity and specificity of a rare diseases for binary outcomes. When the diagnostic test is continuous, under the MAR assumption, Alonzo and Pepe [10] considered the empirical estimate of an ROC curve. They extended some existing imputation and re-weighting methods designed for discrete diagnostic tests to the continuous ones in estimating TPR and FPR. Based on the empirical ROC curve estimate, and by using the trapezoidal rule for integration [11], the estimate of the AUC can be obtained. Their verification bias-corrected estimators are dependent on the modeling of the probability of verification status, and disease status given the screening test result and covariates. A different direction of research deals with nonignorable missing mechanism, which assumes a specific model for the probability of the verification of true disease status, depending on the true disease status. Some early work was done by Zhou [12], [13] and Kosinski and Barnhart [14]. Recently, Rotnitzky, Faraggi, and Schisterman [15] suggested a doubly robust AUC estimator. Fluss et al. [16] extended Rotnitzky et al. [15]'s method, and obtained the asymptotic properties of their estimator.

All of the existing methods differ in modeling the mechanism of the missingness. Throughout the literature on estimation of ROC curves for continuous diagnostic variables, the most popular semi-parametric model of ROC curve assumes binormality. In the binormal model, the diagnostic test variables of non-diseased and diseased groups are normally distributed after some monotone increasing transformation H. Currently, there is no verification bias-corrected method available to estimate the ROC curve under the binormality assumption.

In the absence of the verification bias, a Bayesian method using a rank-based likelihood (BRL) was introduced by Gu and Ghosal [17]. Exploiting the invariance of the rank-likelihood with respect to monotone transformations, they eliminated the need of introducing a prior distribution on the underlying monotone transformation H in the binormal model. They developed Gibbs sampling techniques to simulate samples from the posterior distribution of the parameters in the binormal model, which are obtained to construct a Bayes estimator.

Our proposed verification bias-corrected estimator of ROC is an appropriate modification of the BRL method in the situation with missing labels. We assume that the probability of having labels is a monotone function of the diagnostic measurement. Then the distribution of the unobserved labels conditional on the observations can be easily computed using the Bayes theorem. Hence the missing labels can be imputed within the Gibbs sampling scheme of the BRL method. Coupled with this additional step, the BRL method is hence extended in this partial gold standard situation, and will be abbreviated as PG-BRL.

The following notations will be used. Let ϕ and Φ stand for the density and cumulative distribution function (CDF) of standard normal distribution, respectively, and $\bar{\Phi} = 1 - \Phi$. We use $\phi_{(\mu,\sigma)}$ to denote the density function of the normal distribution $\mathrm{N}(\mu, \sigma^2)$ with mean μ and standard deviation σ. Let $\mathrm{TN}(\mu, \sigma^2, (e_1, e_2))$ denote $\mathrm{N}(\mu, \sigma^2)$ distribution truncated to the interval (e_1, e_2), where $e_1 < e_2$, $e_1, e_2 \in \mathbb{R} \cup \{-\infty, \infty\}$.

The paper is organized as follows. The methodology is described in Section 2. In Section 3, we obtain posterior consistency. Simulation studies and real data analysis are provided in Sections 4 and 5, respectively. Estimation with covariates, and with nonignorable missing mechanism is also discussed in Section 6.

2. Description of the methodology

2.1. Notation

Let $\boldsymbol{S} = \boldsymbol{S}_N = (S_1, \ldots, S_N) = (\boldsymbol{X}, \boldsymbol{Y})$ be the diagnostic measurements associated with N subjects under study, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined in this section below. We denote the number of observations from healthy and diseased groups by m and n respectively, $m+n = N$. Let $D_1, \ldots, D_N$ stand for the true disease status of subjects, where 0 means healthy and 1 means disease. Thus $m = \sum_{i=1}^{N} \mathbb{1}(D_i = 0)$, and $n = \sum_{i=1}^{N} \mathbb{1}(D_i = 1)$, where $\mathbb{1}$ stands for the indicator function.

Under the partial gold standard scenario, we only observe a small fraction of subjects having the true disease status D_i, $i = 1, \ldots, N$. Let $\boldsymbol{L} = (L_1, \ldots, L_N)$, where L_i is defined by

$$L_i = \begin{cases} 0, & \text{if label is observed and } D_i = 0, \\ 1, & \text{if label is observed and } D_i = 1, \\ 2, & \text{if label is not observed.} \end{cases} \tag{1}$$

Observe that in this representation, one single variable carries information on missingness, as well as the labels if observed.

Let m^* and n^* stand for the number of observations having labels from healthy and diseased groups, respectively, i.e., $m^* = \sum_{i=1}^{N} \mathbb{1}(L_i = 0)$, $n^* = \sum_{i=1}^{N} \mathbf{1}(L_i = 1)$, and put $N^* = m^* + n^*$. Let $\boldsymbol{X} = \boldsymbol{X}_m = (X_1, \ldots, X_m) = (S_i : D_i = 0, i = 1, \ldots, N)$ and $\boldsymbol{Y} = \boldsymbol{Y}_n = (Y_1, \ldots, Y_n) = (S_i : D_i = 1, i = 1, \ldots, N)$. Note that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not observable. Let H be the unknown monotone increasing transformation making the transformed observations normally distributed as described in (2) below. Let the transformed measurements denoted by $\boldsymbol{Z} = H(\boldsymbol{X})$, $\boldsymbol{W} = H(\boldsymbol{Y})$ and $\boldsymbol{Q} = H(\boldsymbol{S})$. Let $\widetilde{\boldsymbol{Q}}$, $\widetilde{\boldsymbol{S}}$, $\widetilde{\boldsymbol{L}}$, and $\widetilde{\boldsymbol{D}}$ stand for the order statistic of $\boldsymbol{Q}$, the order statistic of $\boldsymbol{S}$, the labeling and the disease status corresponding to $\widetilde{\boldsymbol{Q}}$, respectively. Moreover, let $\widetilde{Q}_k$ and $\widetilde{S}_k$, denote the kth element $\widetilde{\boldsymbol{Q}}$ and $\widetilde{\boldsymbol{S}}$, respectively.

Let the rank of $\boldsymbol{S}$ be $\boldsymbol{R}_N = R(\boldsymbol{S}) = (R(S_1), \ldots, R(S_N)) = (R_{N1}, \ldots, R_{NN})$. Define the collection of unobserved and observed labels as $\boldsymbol{D}_{\text{un}} = \{D_i : L_i = 2, i \leq N\}$ and $\boldsymbol{D}_{\text{obs}} = \{D_i : L_i = 0 \text{ or } 1, i \leq N\}$, respectively.

2.2. Model

We assume the disease prevalence rate in the population is $0 < \lambda < 1$, i.e., the underlying true disease labels for N subjects as $D_i \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, \lambda)$. Conditional on the labels, we have $S_i \mid \{D_i = 0\} \overset{\text{i.i.d.}}{\sim} F$, $S_i \mid \{D_i = 1\} \overset{\text{i.i.d.}}{\sim} G$, where F and G are continuous CDFs. We assume the binomality assumption holds, i.e., there exists some strictly monotone increasing transformation H, such that

$$Q_i \mid \{D_i = 0\} \overset{\text{i.i.d.}}{\sim} \text{N}(0,1);\ Q_i \mid \{D_i = 1\} \overset{\text{i.i.d.}}{\sim} \text{N}(\mu, \sigma^2),\ \mu > 0, \tag{2}$$

where $Q_i = H(S_i)$.

It is well known that the ROC curve under binormality is given by

$$R(t) = \Phi(a + b\Phi^{-1}(t)),\ \text{where } a = \mu/\sigma,\ b = 1/\sigma. \tag{3}$$

The area under the ROC curve (AUC) also has an explicit form

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right). \tag{4}$$

Typically in a clinical practice, doctors will prescribe more accurate diagnostic test to the patients only if their screening test results show high risk of disease of interest. In particular, subject with a higher diagnostic result will have higher chance of being forwarded to a more thorough gold standard test. Hence, missing the gold standard test completely at random is not an appropriate assumption. In general, we can model the probability of verifying the disease status as

$$\text{P}(L_i \neq 2 \mid Q_i, D_i) = g(Q_i), \tag{5}$$

where g is a monotone increasing function. Note that Q_i's are masked by the unknown transformation H. Hence Q_i's are actually not observed.

Alonzo and Pepe [10] specified the following model based on observed S:

$$P(L_i \neq 2 \mid S) = \begin{cases} 1, & \text{if } S > S_{(p_1 N)}, \\ p_2, & \text{otherwise}, \end{cases} \tag{6}$$

Here, p_1 and p_2 are probabilities known from data source. Since Q's and S's have the same ordering, (6) can be rewritten as a special case of (5) with

$$g(Q) = \begin{cases} 1, & \text{if } Q > Q_{(p_1 N)}, \\ p_2, & \text{if } Q \leq Q_{(p_1 N)}. \end{cases}$$

Another reasonable model uses the probit link:

$$\text{P}(L_i \neq 2 \mid Q_i) = \Phi(\alpha + \beta Q_i),\ \alpha, \beta \text{ unknown and } \beta > 0. \tag{7}$$

2.3. Prior distribution

We will follow a Bayesian approach to estimate (λ, μ, σ), equivalently, (λ, a, b). The prior distributions are described as follows:

- The disease prevalence $\lambda \sim \text{Beta}(l_1, l_0)$, where l_1 and l_0 are chosen to match the mean and the standard error from our prior knowledge, i.e., l_1 and l_0 are chosen to equate to $l_1/(l_1 + l_0)$ with the prior guess for the population prevalence of disease and $\sqrt{l_0 l_1}/((l_0 + l_1)\sqrt{l_0 + l_1 + 1})$ with the anticipated uncertainty in the prior guess.
- In general, it is difficult to specify a subjective prior for (μ, σ). We choose the most commonly used improper prior $\pi(\mu, \sigma) \propto \sigma^{-1}$ for location-scale parameters.

2.4. Posterior distribution

2.4.1. Invariant set. Under the binormality assumption (2), the ranks and labels of $\boldsymbol{Q}$, denoted by $R(\boldsymbol{Q})$ and $L(\boldsymbol{Q})$ respectively, are invariant after the transformation H on $\boldsymbol{S}$. Therefore, a set $\mathcal{D}_{\text{obs}}$ invariant under the action of H can be defined as follows [17]:

$$\begin{aligned} \mathcal{D}_{\text{obs}} &= \{(\boldsymbol{z}, \boldsymbol{w}) \in \mathbb{R}^{m+n} : R(\boldsymbol{z}, \boldsymbol{w}) = R(\boldsymbol{S}),\ L(\boldsymbol{z}, \boldsymbol{w}) = L(\boldsymbol{S})\}, \\ &= \{(\boldsymbol{z}, \boldsymbol{w}) \in \mathbb{R}^{m+n} : \underline{z}_k < z_k < \overline{z}_k,\ \underline{w}_l < w_l < \overline{w}_l\ \forall k, l,\ L(\boldsymbol{z}, \boldsymbol{w}) = L(\boldsymbol{S})\}, \end{aligned} \tag{8}$$

where $\boldsymbol{z} = (z_1, \ldots, z_m)$, $\boldsymbol{w} = (w_1, \ldots, w_n)$, and

$$\begin{aligned} \underline{z}_k &= \max_i\{z_i : R_{Ni} < R_{Nk}\} \vee \max_j\{w_j : R_{N(m+j)} < R_{Nk}\}, \\ \overline{z}_k &= \min_i\{z_i : R_{Nk} < R_{Ni}\} \wedge \min_j\{w_j : R_{Nk} < R_{N(m+j)}\}, \\ \underline{w}_l &= \max_i\{z_i : R_{Ni} < R_{N(m+l)}\} \vee \max_j\{w_j : R_{N(m+j)} < R_{N(m+l)}\}, \\ \overline{w}_l &= \min_i\{z_i : R_{N(m+l)} < R_{Ni}\} \wedge \min_j\{w_j : R_{N(m+l)} < R_{N(m+j)}\}, \end{aligned}$$

for all $k, i = 1, \ldots, m$; $l, j = 1, \ldots, n$. The maximum over the empty set is set to $-\infty$ and the minimum over the empty set is set to ∞.

2.4.2. Lemmas. In order to obtain the posterior distribution and posterior consistency, the following lemmas are needed. The proofs are deferred to the Appendix.

Lemma 1. *For any $c_1 \in \mathbb{R}$ and $c_2 \geq 0$,*

$$\int \bar{\Phi}(c_1 + c_2 t)\phi_{(\mu,\sigma)} dt = \bar{\Phi}\left(\frac{c_1 + c_2\mu}{\sqrt{1 + c_2^2\sigma^2}}\right). \tag{9}$$

Lemma 2. *Let $S_1, \ldots, S_N$ be the independent diagnostic variables with underlying disease status $D_1, \ldots, D_N$ which follow Bin$(1, \lambda)$. Assume that (2) and (5) hold. Then, we have*

$$\text{P}(D_i = 1 \mid Q_i = t, L_i = 2) = \text{P}(D_i = 1 \mid Q_i = t, L_i \neq 2) = \frac{\lambda\phi_{(\mu,\sigma)}(t)}{\lambda\phi_{(\mu,\sigma)}(t) + (1 - \lambda)\phi(t)}, \tag{10}$$

where L_i is defined in (1).

Remark 1. Lemma 2 implies that D and $\mathbb{1}\{L \neq 2\}$ are independent given the value of Q_i. This independence eliminates dependence on the parameters in the model of verification, and makes the calculation much simpler. Further, the expressions are free of $g(\cdot)$, thus allowing us to compute the posterior without actually knowing the verification probability function $g(\cdot)$, as long as it is monotone increasing. In particular, this means that the method is completely protected against misspecification of the verification probability function, which is hard to specify in practice. This is an extremely desirable robustness property of the proposed method.

The following lemma will be used to study the large sample behavior of the posterior distribution, but has no role in computing the posterior given a data set.

Lemma 3. *Assume that (2) and (5) hold. Then the conditional density of Q is given by*

$$f_Q(t \mid L \neq 2) = (1 - \lambda^*_{(\mu,\sigma,g)})\phi^*_{(g)}(t) + \lambda^*_{(\mu,\sigma,g)}\phi^*_{(\mu,\sigma,g)}(t), \tag{11}$$

where $\lambda^*_{(\mu,\sigma,g)}$ and $\phi^*_{(\mu,\sigma,g)}(t)$ are defined as

$$\lambda^*_{(\mu,\sigma,g)} = \frac{\lambda \int g(s)\phi_{(\mu,\sigma)}(s)ds}{\int g(s)\{(1-\lambda)\phi(s) + \lambda\phi_{(\mu,\sigma)}(s)\}ds}, \tag{12}$$

$$\phi^*_{(\mu,\sigma,g)}(t) = \frac{g(t)\phi_{(\mu,\sigma)}(t)}{\int g(s)\phi_{(\mu,\sigma)}(s)ds}, \tag{13}$$

and $\phi^*_{(g)}(t)$ stands for $\phi^*_{(0,1,g)}(t)$.

Remark 2. When (7) holds, (12) and (13) can be simplified to

$$\lambda^*_{(\mu,\sigma,g)} = \frac{\lambda\Phi(\frac{\alpha+\beta\mu}{\sqrt{1+(\beta\sigma)^2}})}{(1-\lambda)\Phi(\frac{\alpha}{\sqrt{1+\beta^2}}) + \lambda\Phi(\frac{\alpha+\beta\mu}{\sqrt{1+(\beta\sigma)^2}})} \tag{14}$$

$$\phi^*_{(\mu,\sigma,g)}(t) = \phi^*_{(\mu,\sigma,\alpha,\beta)}(t) = \frac{\Phi(\alpha+\beta t)\phi_{(\mu,\sigma)}(t)}{\Phi(\frac{\alpha+\beta\mu}{\sqrt{1+(\beta\sigma)^2}})}. \tag{15}$$

2.4.3. Likelihood and posterior distribution. Based on the invariant set (8), a rank-based partial likelihood under the gold standard can be constructed from

$$\mathrm{P}\{(\boldsymbol{Z},\boldsymbol{W}) \in \mathcal{D}_{\mathrm{obs}} | \mu,\sigma\} = \mathrm{P}\{(\boldsymbol{z},\boldsymbol{w}) \in \mathbb{R}^{m+n} : R(\boldsymbol{z},\boldsymbol{w}) = R(\boldsymbol{S}),\ L(\boldsymbol{z},\boldsymbol{w}) = L(\boldsymbol{S}) | \mu,\sigma\}. \tag{16}$$

In the presence of the verification bias, the posterior distribution of (λ,μ,σ), given the ranks and the observed labels, is complicated. However, by applying a data augmentation technique, we can implement Gibbs sampling to compute the posterior distribution. More specifically, augmentation variables $\widetilde{\boldsymbol{Q}}, \boldsymbol{D}_{\mathrm{un}}$ will be used. The resulting posterior distributions of one parameter conditional on the rest of the parameters, ranks and observed labels are relatively simple, and can be described as follows:

- Posterior distribution of (μ,σ) given the rest:

$$\begin{aligned} \sigma^2 | \mathrm{rest} &\sim \text{inverse gamma}((n'-1)/2, (n'-1)s_w^2/2), \\ \mu | \mathrm{rest} &\sim \mathrm{TN}(\bar{W}_{n'}, \sigma^2/n', (0,\infty)), \end{aligned} \tag{17}$$

where $n' = \sum_{i=1}^{N}\{\mathbb{1}(\widetilde{D}_i = 1, \widetilde{L}_i = 2) + \mathbb{1}(\widetilde{L}_i = 1)\}$, $s_w^2 = \sum_{j=1}^{n'}(W_j - \bar{W}_{n'})^2/(n'-1)$, and $\bar{W}_{n'} = \sum_{j=1}^{n'} W_j/n'$.

- $\widetilde{\boldsymbol{Q}}$ can be sequentially updated conditional on (μ, σ) by

$$\widetilde{Q}_i^{(\text{new})}|\text{rest} \sim \begin{cases} \text{TN}(0, 1, (\widetilde{Q}_{i-1}^{(\text{new})}, \widetilde{Q}_{i+1})), & \text{if } \widetilde{D}_i = 0, \\ \text{TN}(\mu, \sigma^2, (\widetilde{Q}_{i-1}^{(\text{new})}, \widetilde{Q}_{i+1})), & \text{if } \widetilde{D}_i = 1, \end{cases} \tag{18}$$

  where $i = 1, \ldots, N$, $\widetilde{Q}_0 = -\infty$, $\widetilde{Q}_{N+1} = \infty$.
- Posterior distribution of λ given the rest:

$$\lambda|\text{rest} \sim \text{Beta}(l_1 + n', l_0 + m'), \tag{19}$$

  where $m' = \sum_{i=1}^{N}\{\mathbb{1}(\widetilde{D}_i = 0, \widetilde{L}_i = 2) + \mathbb{1}(\widetilde{L}_i = 0)\}$.
- Update the augmentation variable $\widetilde{D}_i \in \boldsymbol{D}_{\text{un}}$ by

$$\widetilde{D}_i^{(\text{new})}|\text{rest} \overset{\text{ind}}{\sim} \text{Bin}\left(1, \frac{\lambda\phi_{(\mu,\sigma)}(\widetilde{Q}_i^{(\text{new})})}{\lambda\phi_{(\mu,\sigma)}(\widetilde{Q}_i^{(\text{new})}) + (1-\lambda)\phi(\widetilde{Q}_i^{(\text{new})})}\right), \tag{20}$$

The posterior mean of (μ, σ) was used as the Bayes estimator in the BRL method. In this paper, the posterior median is used in place of posterior mean, primarily because the posterior distribution is often considerably skewed.

2.5. Computational algorithm

By applying a data augmentation technique, we can implement Gibbs sampling to obtain posterior median of $\pi^*(\widetilde{\boldsymbol{Q}}, \mu, \sigma, \lambda, \boldsymbol{D}_{\text{un}}|\boldsymbol{S}, \boldsymbol{L})$. Gibbs sampling procedure can be described as follows:

1. Choose an initial value of (μ, σ). Generate any initial value of $(Q_1, \ldots, Q_N)$ which lies in $\boldsymbol{D}_{\text{obs}}$, in particular, a sample from the product m^* many N(0,1) and n^* many N(μ, σ), where m^* and n^* are the number of observations having labels from healthy and diseased groups, $N^* = m^* + n^*$, and restricting the product measure in $\boldsymbol{D}_{\text{obs}}$. Then, initialize missing labels by a $(N - N^*)$ independent Bin$(1, p_0)$ variables, where $p_0 = l_1/(l_1 + l_0)$.
2. Start the iterations:
   (a) Conditional on (μ, σ), update $\widetilde{\boldsymbol{Q}}$, denoted as $\widetilde{\boldsymbol{Q}}^{(\text{new})}$, with constraint $(\boldsymbol{Z}, \boldsymbol{W}) \in \mathcal{D}_{\text{obs}}$ by following (18).
   (b) Update $\boldsymbol{Z}$ and $\boldsymbol{W}$ values based on $\widetilde{\boldsymbol{Q}}^{(\text{new})}$ and $\widetilde{\boldsymbol{D}}$.
   (c) By following the posterior distributions specified in Section 2.4.3, update $(\mu, \sigma, \lambda, \boldsymbol{D}_{\text{un}})$.
3. After the burn-in period, we obtain the estimates $\hat{a}$ and $\hat{b}$ of the intercept a and the slope b in (3), respectively, by picking up the median of the sampled values of μ/σ and $1/\sigma$, respectively. We can also calculate the $100(1-\gamma)\%$ credible interval for a as $(q_{a,\gamma/2}, q_{a,1-\gamma/2})$, where $q_{a,\gamma/2}$ and $q_{a,1-\gamma/2}$ denote $\gamma/2$ and $1-\gamma/2$ quantiles of the sampled values of a. The credible interval of $(q_{b,\gamma/2}, q_{b,1-\gamma/2})$ for b is similarly defined. We also compute the estimate of AUC and its $100(1-\gamma)\%$ credible interval.

3. Consistency of the posterior

Let (μ_0, σ_0) be the true value of (μ, σ). Consider a joint prior density π on (μ, σ) with respect to the Lebesgue measure (denoted by ν). We shall show the posterior distribution $\Pi(\cdot|R(\boldsymbol{S}), \boldsymbol{L})$ for (μ, σ) is consistent at (μ_0, σ_0), i.e., the posterior distribution concentrates around (μ_0, σ_0) as the sample size increases to infinity; see [18].

Theorem 1. *Assume (2) and (5) hold, where the function g is a monotone increasing functions from $\mathbb{R}$ to $(0, 1)$, and that $\pi(\mu, \sigma) > 0$ a.e. $[\nu]$ over $\mathbb{R}\times\mathbb{R}^+$. Then for any neighborhood $\mathcal{U}_0$ of (μ_0, σ_0), we have that*

$$\lim_{N\to\infty} \Pi((\mu, \sigma) \in \mathcal{U}_0|R(\boldsymbol{S}), \boldsymbol{L}) = 1 \quad a.s. \quad [P^{\infty}_{\mu_0,\sigma_0,g,H}], \tag{21}$$

for (μ_0, σ_0) a.e. $[\nu]$, where $P^{\infty}_{\mu_0,\sigma_0,g,H}$ denotes the true joint distribution of all observations described by (2) and (5).

Since imputation of missing labels is the only additional step in the PG-BRL method compared to the BRL method, we shall first sketch the proof of the consistency of the posterior distribution of the binormal parameters (μ, σ) for the BRL method, and then point out the difference with the present PG-BRL situation.

The main idea behind the proof of posterior consistency for the BRL or PG-BRL method is based on a very general posterior consistency theorem due to Doob. This theorem applies to any type of data, provided it can be shown that the parameter is expressible as a function of the whole sequence of observations (see Theorem 2 in the Appendix). For a rank based method like BRL, this was shown by essentially two major steps in Gu and Ghosal [17]. First, one shows that the "quantile of an observation in the whole mixed population" U_i, as defined in equation (5) of Gu and Ghosal [17], is retrievable from the information of only the relative ranks and the labels at each stage of the asymptotics. Then one finds a consistent estimator of the binormal parameters based on these U_j's and the label information. This shows that the binormal parameters can be expressed as functions of ranks and the labels corresponding to all stages. Now in the PG-BRL method, we need to work with observations with verified labels only, which makes the population distribution different from that of all observations. This leads to a different notion of "population quantile of an observation" U_j as defined by (25) in the Appendix. These U_j's can be retrieved from all ranks and verified labels information in the same way as before. However, the distribution of the U_j's given the labels are different in this situation, and will depend on the verification function g in (5). This would be possible whenever the family of distributions of U_j corresponding to verified disease status is identifiable. The identifiability condition will hold if no two functions in the family $\mathcal{G}$ are such that their ratio is the exponential of a quadratic function. The condition is very mild — it is satisfied by (6), (7) and most conceivable verification mechanisms.

4. Simulation studies

Lemma 2 implies that within an MCMC iteration, updating of $\boldsymbol{D}_{\text{un}}$ will depend on $(\lambda, \mu, \sigma, \boldsymbol{Q})$ only. Thus, none of the parameters in model (5) matter in the MCMC scheme, and hence they could be ignored from a computational point of view. This makes our PG-BRL estimate much

simpler to calculate without a specified form of g in model (5). In this section, we intend to show that verification bias-corrected estimators have less bias and variability than, or are comparable with other well-known methods called semiparametric efficient estimator using the estimated verification probabilities (SPE-E), SPE-E curve adjusted by isotonic regression (SPE-E-A), full imputation (FI), mean score imputation (MSI), and inverse probability weighting (IPW) estimators proposed by Alonzo and Pepe [10].

4.1. Under the bionormality assumption

Sets of simulations letting true AUC equal to 0.65, 0.75, 0.85, respectively, are conducted to compare their estimates of the AUC under different forms of g in model (5), different sample sizes based on 1000 simulated data sets. Within each simulated data setting, two missing mechanisms based on verification models (7) for probit regression and (6) for threshold value, are used to generate the partial gold standard data. In the former case, we fix $\beta = 0.3$, and use the values $\alpha = -0.43$ for AUC=0.65, $\alpha = -0.48$ for AUC=0.75, $\alpha = -0.53$ for AUC=0.85. In the latter case, and $p_1 = 0.8$ and $p_2 = 0.2$ in (6). All these given on an average 36% of the subjects having the labels. The PG-BRL estimates are obtained by 100000 Gibbs samples after burn-in at 5000 for each replication. For each replication, there are total 100 and 200 subjects whose diagnostic test results are generated from $\mathrm{N}(0,1)$ and $\mathrm{N}(\sqrt{1+\sigma^2}\,\Phi^{-1}(\mathrm{AUC}), \sigma^2)$ (see (4)) with a disease prevalence rate 0.25, where σ is equal to 1.5 for all cases. Thus for AUC=0.85, the true values of a and b are 1.2457 and 0.6667, respectively; for AUC=0.75, the true values of a and b are 0.8107 and 0.6667, respectively; for AUC=0.65, the true values of a and b are 0.4631 and 0.6667, respectively. We compare our proposed PG-BRL method with SPE-E, PI, MSI and IPW methods in terms of accuracy.

From Table 1, where the data labels are generated by (7) and (6), respectively, we can see that PG-BRL performs consistently better in term of accuracy in most of the cases when the true AUC is moderate among these 5 methods. When the AUC is equal to 0.75, which is often considered to be a very realistic value in practical situations, PG-BRL clearly is more accuracy than the other methods. However, when AUC goes down to 0.65, SPE-E, SPE-E-A, and IPW methods are less biased than PG-BRL, PI and MSI methods. Comparing mean squared error (MSE), PG-BRL method does have the smallest value among all methods. The estimates of (a, b), which are available only for the PG-BRL method, are shown in Table 2. The accuracy of PG-BRL estimates in Table 2 varies for different verification functions but improves with more samples. Theorem 1 implies accuracy of estimates with sufficient amount of data under various verification functions.

[Insert Table 1 here.]

[Insert Table 2 here.]

4.2. Departure from the bionormality assumption

Since our method is based on the binormality assumption, it is important to study the effect of departure from binormality. We generate $(X_1, \ldots, X_m)$ independently from beta distribution with mean 0.15 and standard deviation 0.15, $(Y_1, \ldots, Y_n)$ independently from beta distribution with mean 0.25 and standard deviation 0.15, where $m = 50$, $n = 150$. The corresponding AUC is equal to 0.715, and 1000 simulated data sets are used in the study. For the parameters in the verification models, we use $\alpha = -0.40$, $\beta = 0.3$ for model (7), and $p_1 = 0.8$, $p_2 = 0.2$ for model (6) to make on an average 36% of the subjects having the labels. The PG-BRL estimates are

obtained by 100000 Gibbs samples after burn-in at 5000 for each replication. From the results shown in Table 3, we can see that even though the PG-BRL estimator has higher bias than SPE-E and IPW estimators, but its MSE is much smaller. The simulation results show that our procedure has reasonable robustness properties against departure from binormality. [Insert Table 3 here.]

4.3. Departure from MAR assumption

When MAR assumption does not hold, the dependence of the verification probability in (5) on disease status will make the probabilities in (10) different and dependent on the functional form of g in (5). More precisely, let

$$\mathrm{P}(L_i \neq 2 \mid Q_i, D_i = 0) = g_0(Q_i), \tag{22}$$

$$\mathrm{P}(L_i \neq 2 \mid Q_i, D_i = 1) = g_1(Q_i), \tag{23}$$

then proceeding as in the proof of Lemma 2, it follows that

$$\mathrm{P}(D_i = 1 \mid Q_i = t, L_i = 2) = \frac{\lambda g_1(t)\phi_{(\mu,\sigma)}(t)}{\lambda g_1(t)\phi_{(\mu,\sigma)}(t) + (1-\lambda)g_0(t)\phi(t)}. \tag{24}$$

Here, we illustrate a case non-MAR (NMAR) where the MAR assumption fails. Assume $\mathrm{P}(D = 1) = 0.25$. To achieve 36% verification, i.e., $\mathrm{P}(L_i \neq 2) = 0.36$, we can choose $\mathrm{P}(L_i \neq 2 \mid D = 1) = 0.80$, and hence will need to set $\mathrm{P}(L_i \neq 2 \mid D = 0) = 0.213$. Let $g_0(t) = \Phi(\alpha_0 + \beta_0 t)$, and $g_1(t) = \Phi(\alpha_1 + \beta_1 t)$. We wish to choose pairs $(\alpha_0, \beta_0) \neq (\alpha_1, \beta_1)$ (i.e., NMAR) which will yield verification probabilities 0.80 and 0.213 under disease and non-disease situations respectively. To this end, use Lemma 1 with $\beta_0 = \beta_1 = 0.3$, and solve the following equations:

$$\Phi\left(\frac{\alpha_1 + \beta_1\mu}{\sqrt{1+\beta_1^2\sigma^2}}\right) = 0.8, \qquad \Phi\left(\frac{\alpha_0}{\sqrt{1+\beta_0^2}}\right) = 0.213,$$

given $\alpha_0 = -0.831$ and $\alpha_1 = 0.558$, when the actual values of μ and σ are 1.216 and 1.5, respectively.

We compare estimates shown in Table 4 by NMAR with MAR imputed by threshold model, where simulation setting for MAR exactly follows Section 4.1 with true AUC equal to 0.75. We observe that even under departure from the MAR assumption, the proposed PG-BRL method leads to smaller MSE compared with other methods. [Insert Table 4 here.]

5. Real data analysis

One common issue with ROC curves and biopsy verification occurs whenever biopsy results are used as the gold standard for other measurements, usually with respect to fibrosis or steatosis. The biopsy may not be representative of the liver as a whole and there is always interpretation noise (inter and intra-observer variability). The clinical test may not be a perfect match for the histological observation. However, to explore the relationship between the specified liver enzyme test and the degree of cholestasis in Drug Induced Liver Injury (DILI), liver biopsy database is still helpful. Here, the specified liver enzyme tests include serum total bilirubin

(STB) level and alkaline phosphatase value (AKP). Because the bilirubin may rise without the accumulation of visible bile in the liver, there is not a perfect match between the serum bilirubin and the biopsy. Nevertheless, the biopsy still gives precise information about the degree of cholestasis.

We shall use the data from Drug Induced Liver Injury Network (DILIN). The DILIN, sponsored by the National Institutes of Health in 2003, is a consortium of a data coordinating center, five academic medical centers later expanded to nine starting from the renewal of the second five-year grant. The DILIN prospective study is an ongoing observational study of over 1000 patients with suspected liver injury due to various drugs and complementary and alternative medications (CAM). The goals of this study include the earliest recognition of DILI, the development of standardized instruments and terminology to help identify cases of DILI, investigating clinical and genetic risk factors that predict DILI. The study design and development of the prospective study, and the process of causality assessment, were presented by Fontana et al. [19], and Rockey et al. [20], respectively.

Up to January 2011, there were 898 subjects enrolled in DILIN prospective study. Among these subjects, we include 405 subjects in this study, who had STB and AKP values at DILI first abnormal date (DILI onset date), and had causality score either definitely, very likely, or probably adjudicated by DILIN causality committee. The subjects follow up on the sixth month after their baseline visits. The biopsy samples were restricted to either baseline or historical biopsy samples collected within 60 days of DILI onset and closest to DILI onset for multiple biopsy data. We dichotomize the degree of cholestasis into 0-1 or 2-3 to reduce intra-observer variability, where 80 subjects (56%) had no disease and 64 subjects (44%) had disease. The STB or AKP values were chosen to be closest to the biopsy date with a time window of 7 days. If not, then the corresponding lab value at onset was used instead. Since only 144 subjects out of 405 had biopsy results, verification bias may occur if the analysis is just based on these 144 subjects with the degree of cholestasis. Hence, verification bias-correction methods are needed in this content. By using PG-BRL and other methods listed in the simulation study, we have the following results for STB and AKP in Table 5 and Figure 1.

[Insert Table 5 here.]

[Insert Figure 1 here.]

We obtain point-estimates of the AUC using the proposed PG-BRL method based on the measurements STB and AKP, which are given by 0.7388 and 0.6550. This seems to indicate that categorization of degree of cholestatis based on STB is more accurate than that based on AKP. A formal comparison of the accuracies of these two diagnostic measurements will have to involve a test of hypothesis of the equality of the two AUC values, which will necessarily involve a joint modeling of both STB and AKP for each subject. This will allow us to draw MCMC sample from the joint posterior distribution of all parameters in the two binormal models for STB and AKP, and consequently from the posterior distribution of the difference or ratio of the corresponding AUC. Presently, such a joint binormal model has not been developed, so we refrain from such an analysis.

Although SPE-E estimator is "doubly robust", ensuring either verification model or disease model can be challenging. A difficulty with the SPE-E method is that the estimated ROC curve may not be monotone due to negative component in the formula of estimates of $\mathrm{TPR}(c)$ and $\mathrm{FPR}(c)$. Hence, SPE-E-A ROC curve is displayed in Figure 1 instead of SPE-E ROC curve.

For all the other approaches we used to compare with our method, it is difficult to know or

estimate $S_{(p_1 N)}$ in (6), since two populations of AKP values are well-mixed. Thus, we conclude that PG-BRL method is more flexible when the disease and non-disease population are well mixed, and no hyper-parameters are needed in its calculation.

6. Discussion

In this paper, we considered modeling the diagnostic variable as an independent variable in the binormal model. Sometimes a set of covariates, say $\boldsymbol{V}$, is also observed, which contributes a linear effect through an unknown regression coefficient $\boldsymbol{\beta}$, i.e., the binormal model will become

$$H(X_i - \boldsymbol{\beta}^T \boldsymbol{V}_i) \mid D_i = 0 \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1),$$
$$H(Y_i - \boldsymbol{\beta}^T \boldsymbol{V}_i) \mid D_i = 1 \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2).$$

In this case, the BRL and the PG-BRL methods are not applicable since the ranks of the covariate adjusted observations may change between MCMC iterations due to the changes in the sampled value of $\boldsymbol{\beta}$. This will destroy the invariance structure that the rank-likelihood is based on. In order to accommodate covariates in the BRL and PG-BRL methods, we can adopt a preprocessing step to replace the diagnostic test values with the residuals after adjusting for the covariates using a regression model and estimating the regression coefficient by the method of least squares. If a handful of covariates among several needs to be selected as well, then the LASSO estimator may be used in place of the least square estimator. Of course, a fully Bayesian analysis is possible by using the full likelihood instead of the rank-likelihood, but that will need a prior on the transformation H as well.

In a clinical practice, sometimes the MAR assumption may be violated. For instance, when a patient is too sick or old, or is unable to bear the cost of the gold standard test, he or she may not be able to follow the physician's recommendation to take the gold standard test. In these cases, probability of verification will depend on covariates and disease status as well unless all the relevant health and financial information are collected also, and are properly accounted in the model (5) for verification. The PG-BRL is nevertheless applicable, but the posterior will depend on the verification functions g_0 and g_1 in (22) and (23), respectively, which will have to be explicitly known.

Appendix

Proof of Lemma 1. The result is well known, but we give a proof for completeness. First, we can show that

$$\int \bar{\Phi}(c_1 + c_2 t)\phi(t)dt = \bar{\Phi}\left(\frac{c_1}{\sqrt{1+c_2^2}}\right).$$

Let $J(c_1) = \int \bar{\Phi}(c_1 + c_2 t)\phi(t)dt$. Then, by differentiation under the integral sign which can be

justified by the dominated convergence theorem,

$$\begin{aligned} J'(c_1) &= -\int \phi(c_1 + c_2 t)\phi(t)dt \\ &= -\int \frac{1}{2\pi} \exp\left\{-\frac{1}{2}\left(c_1^2 + 2c_1c_2t + c_2^2t^2 + t^2\right)\right\} dt \\ &= -\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{c_1^2}{2(1+c_2^2)}\right\} \int \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(u+w)^2\right\} \frac{du}{\sqrt{1+c_2^2}} \\ &= -\phi\left(\frac{c_1}{\sqrt{1+c_2^2}}\right) \frac{1}{\sqrt{1+c_2^2}}, \end{aligned}$$

where we have used the substitution $u = t\sqrt{1+c_2^2}$, $w = c_1c_2/\sqrt{1+c_2^2}$. Hence, $J(c_1) = -\Phi(c_1/\sqrt{1+c_2^2}) + c$, for some constant c. Since $J(\infty) = 0$, we have $c = 1$, and hence $J(c_1) = 1 - \Phi(c_1/\sqrt{1+c_2^2}) = \bar{\Phi}(c_1/\sqrt{1+c_2^2})$.

Now letting $s = \frac{t-\mu}{\sigma}$, we have

$$\int \bar{\Phi}(c_1 + c_2 t)\phi_{(\mu,\sigma)}dt = \int \bar{\Phi}(c_1 + c_2\mu + c_2\sigma s)\phi_(s)ds = \bar{\Phi}\left(\frac{c_1 + c_2\mu}{\sqrt{1+c_2^2\sigma^2}}\right).$$

Proof of Lemma 2. By Bayes' theorem, we have

$$\begin{aligned} &\mathrm{P}(D_i = 1 \mid Q_i = t, L_i = 2) \\ &= \frac{\mathrm{P}(D_i = 1) f_{Q_i}(t \mid D_i = 1)\mathrm{P}(L_i = 2 \mid Q_i = t, D_i = 1)}{\sum_{d=0}^{1}\{\mathrm{P}(D_i = d) f_{Q_i}(t \mid D_i = d)\mathrm{P}(L_i = 2 \mid Q_i = t, D_i = d)\}} \end{aligned}$$

Since $f_{Q_i}(t \mid D_i = 1, L_i = 2) = f_{Q_i}(t \mid D_i = 1)$ and $\mathrm{P}(L_i = 2 \mid Q_i = t, D_i = 1) = \mathrm{P}(L_i = 2 \mid Q_i = t, D_i = 0) = 1 - g(t)$, we have

$$\begin{aligned} \mathrm{P}(D_i = 1 \mid Q_i = t, L_i = 2) &= \frac{\lambda\phi_{(\mu,\sigma)}(t)(1-g(t))}{\lambda\phi_{(\mu,\sigma)}(t)(1-g(t)) + (1-\lambda)\phi(t)(1-g(t))} \\ &= \frac{\lambda\phi_{(\mu,\sigma)}(t)}{\lambda\phi_{(\mu,\sigma)}(t) + (1-\lambda)\phi(t)}, \end{aligned}$$

and $\mathrm{P}(L_i = 2) = \int \mathrm{P}(L_i = 2 \mid Q_i = t) f_{Q_i}(t)dt = \int (1-g(t))\{\lambda\phi_{(\mu,\sigma)}(t) + (1-\lambda)\phi(t)\}dt$.

By following the same lines, we have

$$\begin{aligned} \mathrm{P}(D_i = 1 \mid Q_i = t, L_i \neq 2) &= \frac{\lambda\phi_{(\mu,\sigma)}(t)g(t)}{\lambda\phi_{(\mu,\sigma)}(t)g(t) + (1-\lambda)\phi(t)g(t)} \\ &= \frac{\lambda\phi_{(\mu,\sigma)}(t)}{\lambda\phi_{(\mu,\sigma)}(t) + (1-\lambda)\phi(t)}, \end{aligned}$$

and $\mathrm{P}(L_i \neq 2) = \int \mathrm{P}(L_i \neq 2 \mid Q_i = t) f_{Q_i}(t)dt = \int g(t)\{\lambda\phi_{(\mu,\sigma)}(t) + (1-\lambda)\phi(t)\}dt$.

Proof of Lemma 3. By Bayes' theorem, we have

$$\begin{aligned} f_Q(t \mid L \neq 2) &= \frac{\mathrm{P}(L=0 \text{ or } 1 \mid Q=t) f_Q(t)}{\int \mathrm{P}(L=0 \text{ or } 1 \mid Q=s) f_Q(s) ds} \\ &= \frac{g(t)\{(1-\lambda)\phi(t) + \lambda\phi_{(\mu,\sigma)}(t)\}}{\int g(s)\{(1-\lambda)\phi(s) + \lambda\phi_{(\mu,\sigma)}(s)\} ds} \\ &= \frac{g(t)\phi(t)}{\int g(s)\phi(s) ds} \times \frac{(1-\lambda)\int g(s)\phi(s) ds}{\int g(s)\{(1-\lambda)\phi(s) + \lambda\phi_{(\mu,\sigma)}(s)\} ds} \\ &\quad + \frac{g(t)\phi_{(\mu,\sigma)}(t)}{\int g(s)\phi_{(\mu,\sigma)}(s) ds} \times \frac{\lambda\int g(s)\phi_{(\mu,\sigma)}(s) ds}{\int g(s)\{(1-\lambda)\phi(s) + \lambda\phi_{(\mu,\sigma)}(s)\} ds} \\ &= (1-\lambda^*_{(\mu,\sigma,g)})\phi^*_{(g)}(t) + \lambda^*_{(\mu,\sigma,g)}\phi^*_{(\mu,\sigma,g)}(t), \end{aligned}$$

where $\lambda^*_{(\mu,\sigma,g)}$ and $\phi^*_{(\mu,\sigma,g)}(t)$ are defined in (12) and (13), respectively.

The following version of a well known theorem by Doob's theorem was used, whose proof can be found in Ghosal and Van der Vaart [21].

Theorem 2 (Doob's Theorem) *Let $X^{(n)}$ be observations whose distribution depends on a parameter θ, and both $X^{(n)}$ and θ take values in Polish spaces. Let Π be a prior distribution on θ. Assume that θ is equivalent to a measurable function f on $(X^{(n)} : n \geq 1)$, i.e., $\theta = f(X^{(n)} : n \geq 1)$ a.s. with respect to the joint distribution of θ and $(X^{(n)} : n \geq 1)$. Then the posterior $\Pi(\cdot|X^{(n)})$ is strongly consistent at θ for almost every θ $[\Pi]$.*

Proof of Theorem 1. Since the posterior distribution of (μ, σ) given the ranks and labels do not depend on g and H, we may treat them as known for the purpose of theoretically studying consistency of the posterior distribution of (μ, σ), even though g and H need not be actually known.

Let Ω_N stand for a set of all permutations of $\{1, \ldots, N\}$. If we can show there exists a function $h^* : \Omega_1 \times \Omega_2 \times \ldots \times \{0,1,2\}^\infty \to \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$, such that $(\mu, \sigma) = h^*(\boldsymbol{R}_N, N \geq 1, (L_1, L_2, \ldots))$, then by applying Doob's Theorem, (21) holds.

Let $1 \leq i_1 < \cdots < i_{N^*} \leq N$ be the collection of indices for which $L_{i_j} = 0$ or 1, and $j = 1, \ldots, N^*$. From Lemma 3, $f_Q(t \mid L_i \neq 2) = (1-\lambda^*_{(\mu,\sigma,g)})\phi^*_{(g)}(t) + \lambda^*_{(\mu,\sigma,g)}\phi^*_{(\mu,\sigma,g)}(t)$, with notations as in Lemma 3. Hence, disregarding the disease status, we can regard the overall sample coming independently from the mixture distribution, that is, $Q_{i_j} \overset{\text{i.i.d.}}{\sim} (1-\lambda^*_{(\mu,\sigma,g)})\Phi^*_{(g)} + \lambda^*_{(\mu,\sigma,g)}\Phi^*_{(\mu,\sigma,g)}$, where $\Phi^*_{(g)}(t)$ and $\Phi^*_{(\mu,\sigma,g)}$ are the CDFs of $\phi^*_{(g)}(t)$ and $\phi^*_{(\mu,\sigma,g)}(t)$, respectively. Thus we have

$$U_j = \{(1-\lambda^*_{(\mu,\sigma,g)})\Phi^*_{(g)} + \lambda^*_{(\mu,\sigma,g)}\Phi^*_{(\mu,\sigma,g)}\}(Q_{i_j}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1). \tag{25}$$

Let $(R'_{N^*1}, \ldots, R'_{N^*N^*})$ be the rank vector of $(U_1, \ldots, U_{N^*})$, and $(L'_{N^*1}, \ldots, L'_{N^*N^*})$ stand for their labels. Now, as in Theorem a on page 157 of Hájek and Šidák [22], we have

$$\begin{aligned} \mathrm{E}\left(U_j - \frac{R'_{N^*i_j}}{N^*+1}\right)^2 &= \frac{1}{N^*}\sum_{k=1}^{N^*} \mathrm{E}\left[\left(U_j - \frac{k}{N^*+1}\right)^2 \middle| R'_{N^*i_j} = k\right] \\ &= \frac{1}{N^*}\sum_{k=1}^{N^*} \frac{k(N^*-k+1)}{(N^*+1)^2(N^*+2)} < \frac{1}{N^*}, \end{aligned}$$

which tends to 0 as $N \to \infty$, where the expectation is interpreted as conditional on the labels. Therefore,

$$U_j = \lim_{k\to\infty} \frac{R'_{N_k^* i_j}}{N_k^* + 1} \tag{26}$$

for $j \geq 1$, with probability 1 for some subsequence $\{N_k^*\}$ of $\{N^*\}$, and hence $U_j = h_j(\boldsymbol{R}_N, N \geq 1, L_1, L_2, \ldots)$ for some function $h_j : \Omega_1 \times \Omega_2 \times \ldots \times \{0,1,2\}^\infty \to [0,1]$.

Now given $\{Q_{i_j} : L_{i_j} = 0\} \overset{\text{i.i.d.}}{\sim} \Phi^*_{(g)}$, so that $\{U_j : L_{i_j} = 0\} \overset{\text{i.i.d.}}{\sim} V_{(\mu,\sigma,g)}$, say, where $V_{(\mu,\sigma,g)}$ is the distribution of $\{(1-\lambda^*_{(\mu,\sigma,g)})\Phi^*_{(g)} + \lambda^*_{(\mu,\sigma,g)}\Phi^*_{(\mu,\sigma,g)}\}(\xi)$, $\xi \sim \Phi^*_{(g)}$. Since clearly $V_{(\mu,\sigma,g)}$ is consistently estimable by the empirical distribution of $(U_j : L_{i_j} = 0)$, it now suffices to show that the family $\{V_{(\mu,\sigma,g)} : \mu, \sigma > 0\}$ is identifiable.

If (μ_1, σ_1) and (μ_2, σ_2) are such that $V_{(\mu_1,\sigma_1,g)} = V_{(\mu_2,\sigma_2,g)}$, then it follows that $\lambda^*_{(\mu_1,\sigma_1,g)} = \lambda^*_{(\mu_2,\sigma_2,g)}$ and $\phi^*_{(\mu_1,\sigma_1)} = \phi^*_{(\mu_2,\sigma_2)}$, which then implies that $(\mu_1, \sigma_1) = (\mu_2, \sigma_2)$.

Therefore, there exists a function h of $(U_1, U_2, \ldots, L_1, L_2, \ldots)$ and hence h^* of all ranks and observed labels such that a.s. $[P^\infty_{\mu_0,\sigma_0,g,H}]$, we have

$$\begin{aligned}(\mu, \sigma, g) &= h(U_1, U_2, \ldots,) \\ &= h(h_1(\boldsymbol{R}_N, N \geq 1, L_1, L_2, \ldots), h_2(\boldsymbol{R}_N, N \geq 1, L_1, L_2, \ldots), \ldots) \\ &= h^*(\boldsymbol{R}_N, N \geq 1, L_1, L_2, \ldots).\end{aligned}$$

ACKNOWLEDGEMENTS

Research of the second author is partially supported by NSF grant number DMS-0349111.

REFERENCES

1. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostics Imaging* 1989; **29**:307–335.
2. Poynard T, Imbert-Bismut F, Ratziu V. Serum markers of liver fibrosis. *Hepatology Reviews* 2004; **1(1)**:23–31.
3. Grant A, Neuberger J. Guidelines on the use of liver biopsy in clinical practice. *British Society of Gastroenterology* 1999; **Gut 45 Suppl 4**:IV1–CIV11.
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* New York: John Wiley and Sons., 1987.
5. Begg CB, Greenes R. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; **39**:207–215.
6. Reilly M, Pepe MS. A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika* 1995; **82**:299–314.
7. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* 1998; **7**:337–353.
8. Clayton D, Spiegelhalter D, Dunn G, Pickles A. Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B* 1998; **60**:71–87.
9. Gastwirth JL, Johnson WO, Reneau DM. Bayesian analysis of screening data: application to AIDS in blood donors. *The Canadian Journal of Statistics* 1991; **19**:135–150.
10. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society, Series C* 2005; **54**:173–190.
11. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
12. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communication in Statistics — Theory and Methods* 1993; **22**:3177–3198.

13. Zhou XH. Effect of verification bias on positive and negative predictive values. *Statistics in Medicine* 1994; **13**:1737–1745.
14. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* 2003; **59**:163–171.
15. Rotnitzky A, Faraggi D, Schisterman E. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association* 2006; **101**:1276–1288.
16. Fluss R, Reiser B, Faraggi D, Rotnitzky A. Estimation of the ROC curve under verification bias. *Biometrical Journal* 2009; **3**:475–490.
17. Gu J, Ghosal S. Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference* 2009; **139(6)**:2076–2083.
18. Ghosh JK, Ramamoorthi RV. *Bayesian Nonparametrics.* New York: Springer-Verlag, 2003.
19. Fontana RJ, Watkins PB, Bonkovsky HL, Chalasani N, Davern T, Serrano J, Rochon J. Drug-induced liver injury network (DILIN) prospective study: rationale, design and conduct. *Drug Safety* 2009; **32**:55–68.
20. Rockey DC, Seeff LB, Rochon J, Freston J, Chalasani N, Bonacini M, Fontana RJ, Hayashi PH. Causality assessment in drug-induced liver injury using a structured expert opinion process: Comparison to the Roussel-Uclaf causality assessment method. *Hepatology* 2010; **51**:2117–2126.
21. Ghosal S, Van der Vaart AW. *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press (to appear), 2015.
22. Hájek, J. and Šidák, Z. *Theory of Rank Tests.* New York: Academic Press, 1967.

Table 1. Simulation studies: Estimates of AUC using methods PG-BRL, SPE-E, SPE-E-A, FI, MSI, IPW.

Imputation		$N = 100$		$N = 200$	
Model	Method	Bias	MSE	Bias	MSE
Probit	PG-BRL	−0.0321	0.0085	−0.0156	0.0046
AUC=.85	SPE-E	0.0162	0.0066	0.0136	0.0036
	SPE-E-A	0.0156	0.0065	0.0135	0.0036
	FI	0.0357	0.0050	0.0349	0.0032
	MSI	0.0309	0.0050	0.0303	0.0031
	IPW	−0.0349	0.0100	−0.0357	0.0060
Probit	PG-BRL	0.0155	0.0068	0.0084	0.0043
AUC=.75	SPE-E	0.0339	0.0126	0.0208	0.0062
	SPE-E-A	0.0335	0.0125	0.0207	0.0062
	FI	0.0624	0.0113	0.0535	0.0065
	MSI	0.0559	0.0111	0.0465	0.0060
	IPW	−0.0188	0.0137	−0.0312	0.0078
Probit	PG-BRL	0.0622	0.0083	0.0529	0.0059
AUC=.65	SPE-E	0.0360	0.0164	0.0343	0.0093
	SPE-E-A	0.0358	0.0164	0.0343	0.0093
	FI	0.0729	0.0168	0.0725	0.0109
	MSI	0.0647	0.0160	0.0644	0.0100
	IPW	−0.0130	0.0154	−0.0117	0.0082
Threshold	PG-BRL	−0.0067	0.0078	0.0013	0.0038
AUC=.85	SPE-E	0.0144	0.0098	0.0110	0.0045
	SPE-E-A	0.0129	0.0096	0.0104	0.0044
	FI	0.0391	0.0060	0.0383	0.0035
	MSI	0.0329	0.0059	0.0330	0.0033
	IPW	0.0068	0.0100	0.0078	0.0046
Threshold	PG-BRL	0.0220	0.0077	0.0146	0.0052
AUC=.75	SPE-E	0.0213	0.0146	0.0120	0.0077
	SPE-E-A	0.0199	0.0143	0.0116	0.0076
	FI	0.0595	0.0107	0.0572	0.0073
	MSI	0.0517	0.0103	0.0493	0.0067
	IPW	0.0098	0.0151	0.0055	0.0078
Threshold	PG-BRL	0.0687	0.0105	0.0482	0.0053
AUC=.65	SPE-E	0.0245	0.0203	0.0182	0.0091
	SPE-E-A	0.0233	0.0199	0.0180	0.0090
	FI	0.0744	0.0167	0.0729	0.0104
	MSI	0.0645	0.0159	0.0634	0.0094
	IPW	0.0114	0.0206	0.0103	0.0090

Table 2. Simulation studies: Estimates of a and b using method PG-BRL.

Imputation		$N = 100$		$N = 200$	
Model	Method	Bias	MSE	Bias	MSE
Probit	PG-BRL a	0.0134	0.4438	0.0121	0.2236
AUC=.85	b	0.0044	0.0846	−0.0080	0.0442
Probit	PG-BRL a	0.1764	0.3247	0.0844	0.1314
AUC=.75	b	0.0206	0.0794	0.0063	0.0331
Probit	PG-BRL a	0.2792	0.2179	0.2209	0.1209
AUC=.65	b	0.0253	0.0714	0.0409	0.0332
Threshold	PG-BRL a	0.2420	0.6750	0.1483	0.3226
AUC=.85	b	0.0955	0.1097	0.0506	0.0535
Threshold	PG-BRL a	0.2409	0.4305	0.1316	0.1905
AUC=.75	b	0.0667	0.0927	0.0310	0.0373
Threshold	PG-BRL a	0.3382	0.3463	0.1995	0.1046
AUC=.65	b	0.0684	0.0833	0.0373	0.0268

Table 3. Simulation studies (Departure from bionormality assumption): Estimates of AUC using methods PG-BRL, SPE-E, SPE-E-A, FI, MSI, IPW.

Imputation		$N = 200$	
Model	Method	Bias	MSE
Probit	PG-BRL	0.0257	0.0022
AUC=0.715	SPE-E	0.0070	0.0040
	SPE-E-A	0.0055	0.0040
	FI	−0.0560	0.0076
	MSI	−0.0381	0.0055
	IPW	0.0066	0.0042
Threshold	PG-BRL	0.0268	0.0026
AUC=.715	SPE-E	0.0063	0.0057
	SPE-E-A	0.0017	0.0052
	FI	−0.0930	0.0125
	MSI	−0.0746	0.0093
	IPW	0.0041	0.0055

Table 4. Simulation studies (Departure from MAR assumption): Estimates of AUC using methods PG-BRL, SPE-E, SPE-E-A, FI, MSI, IPW.

Assumption		$N = 200$	
Model	Method	Bias(True AUC=0.75)	MSE
MAR	PG-BRL	0.0146	0.0052
Threshold	SPE-E	0.0120	0.0077
	SPE-E-A	0.0116	0.0076
	FI	0.0572	0.0073
	MSI	0.0493	0.0067
	IPW	0.0055	0.0078
NMAR	PG-BRL	−0.0569	0.0052
	SPE-E	−0.0855	0.0126
	SPE-E-A	−0.0855	0.0126
	FI	−0.0408	0.0050
	MSI	−0.0479	0.0058
	IPW	−0.1184	0.0188

Table 5. Real data setting: Estimates of AUCs and corresponding 95% confidence intervals (CI) for Total Bilirubin and Alkaline phosphatase.

Lab Test	Method	$\widehat{AUC}(\hat{sd})$	95% CI
STB	PG-BRL	0.7388(0.0559)	(0.6389,0.8601)
	SPE-E	0.7198(0.0655)	(0.5970,0.8495)
	SPE-E-A	0.7109(0.0573)	(0.5938,0.8145)
	FI	0.6704(0.0357)	(0.6055, 0.7337)
	MSI	0.7041(0.0417)	(0.6201,0.7849)
	IPW	0.6714(0.0481)	(0.5790,0.7643)
AKP	PG-BRL	0.6550(0.0738)	(0.5624,0.8527)
	SPE-E	0.6283(0.0962)	(0.4421,0.8183)
	SPE-E-A	0.6248(0.0865)	(0.4517,0.7727)
	FI	0.5683(0.0289)	(0.5156,0.6353)
	MSI	0.5953(0.0447)	(0.5036,0.6788)
	IPW	0.5768(0.0498)	(0.4693,0.6716)

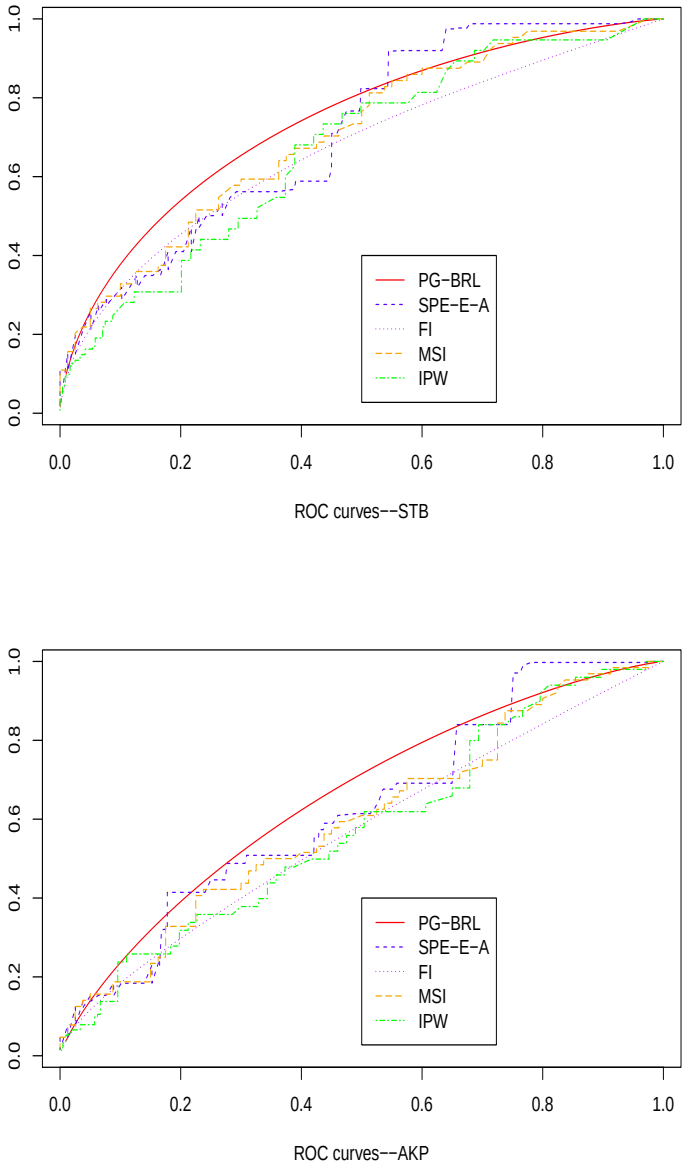

Figure 1. Real data analysis: ROC curve estimates for STB and AKP.