

## Bayesian ROC curve estimation under verification bias

Jiezhun Gu<sup>1,\*</sup>, Subhashis Ghosal<sup>2,\*\*</sup>, and David E. Kleiner<sup>3,\*\*\*</sup>

<sup>1</sup>Duke Clinical Research Institute, Duke University Medical Center, Durham, NC 27715, USA

<sup>2</sup>Department of Statistics, North Carolina State University, Raleigh, NC, 27695, USA

<sup>3</sup>Laboratory of Pathology, National Cancer Institute, Bethesda, MD, 20892, USA

\**email*: jiezhun.gu@duke.edu

\*\**email*: subhashis.ghosal@ncsu.edu

\*\*\**email*: kleinerd@mail.nih.gov

**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 could be invasive, expensive, and its result is not always available for the whole population. 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.

**KEY WORDS:** Binormal model; MAR assumption; Posterior consistency; ROC curve; Verification bias-correction.

## 1. Introduction

The Receiver Operating Characteristic (ROC) curve has long been widely used in diagnostic medicine (Hanley, 1989) because of its ability to incorporate all the decision information 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 a 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.

Since gold standard tests could 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 costs over US \$1000 without complications and about US \$3000 with complications (Poynard, Imbert-Bismut, and Ratziu, 2004). 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 (Grant and Neuberger, 1999).

Because of a differential in the chance of verification between subjects with high and low risk, it follows that the verified subjects are not a random sample 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. The commonly used assumption for missing verification of disease status assumes missing at

random (MAR) introduced by Little and Rubin (1987), which means the chance of missing the verification of disease status is independent of the disease itself conditional on the observed measurements. Below we argue why MAR assumption is reasonable in our context. The decision for a physician to order a gold standard test is primarily based on a patient's screening test results, signs and symptoms, medical history, current health condition, and patient's consent to withstand the invasive gold standard procedure. Hence, it is reasonable to assume that, conditional on these factors, the true disease status does not affect the probability of verification.

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 (1983) proposed an adjusted estimate of TPR and FPR. Other more efficient methods are discussed by Reilly and Pepe (1995), Zhou (1998), Clayton et al. (1998). When the diagnostic test is continuous, under the MAR assumption, Alonzo and Pepe (2005) 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 (Bamber, 1975), 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 (1993, 1994), Kosinski and Huiman (2003). Recently, Rotnitzky, Faraggi, and Schisterman (2006) suggested a doubly robust AUC estimator. Fluss et al. (2009) extended Rotnitzky et al. (2006)'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 (2009). Exploiting the invariance of the rank-likelihood with respect to monotone transformation, 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. In addition to binormality, we assume that the probability of having labels is a monotone function of the diagnostic measurement only, and hence in particular, the MAR assumption holds. 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 the partial gold standard situation, and will be abbreviated as PG-BRL.

The following notations will be used. Let  $\phi$  and  $\Phi$  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  $N(\mu, \sigma^2)$  with mean

$\mu$  and standard deviation  $\sigma$ . Let  $\text{TN}(\mu, \sigma^2, (e_1, e_2))$  denote  $N(\mu, \sigma^2)$  distribution truncated to the interval  $(e_1, e_2)$ , where  $e_1 < e_2$ ,  $e_1, e_2 \in \mathbb{R}$ .

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. We conclude with a discussion in Section 6.

## 2. Description of the methodology

### 2.1 Notation

Let  $\mathbf{S} = \mathbf{S}_N = (S_1, \dots, S_N) = (\mathbf{X}, \mathbf{Y})$  be the diagnostic measurements associated with  $N$  subjects under study, where  $\mathbf{X}$  and  $\mathbf{Y}$  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  $D_1, \dots, D_N$  stand for the true disease status of subjects, where 0 means healthy and 1 means disease. Thus  $m = \sum_{i=1}^N 1(D_i = 0)$ , and  $n = \sum_{i=1}^N 1(D_i = 1)$ .

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

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

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.,  $m^* = \sum_{i=1}^N 1(L_i = 0)$ ,  $n^* = \sum_{i=1}^N 1(L_i = 1)$ , and put  $N^* = m^* + n^*$ . Let  $\mathbf{X} = \mathbf{X}_m = (X_1, \dots, X_m) = \{S_i : D_i = 0, i = 1, \dots, N\}$  and  $\mathbf{Y} = \mathbf{Y}_n = (Y_1, \dots, Y_n) = \{S_i : D_i = 1, i = 1, \dots, N\}$ . 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  $\mathbf{Z} = H(\mathbf{X})$ ,  $\mathbf{W} = H(\mathbf{Y})$  and  $\mathbf{Q} = H(\mathbf{S})$ . Let  $\tilde{\mathbf{Q}}$ ,  $\tilde{\mathbf{S}}$ ,  $\tilde{\mathbf{L}}$ , and  $\tilde{\mathbf{D}}$  stand for the order statistic of  $\mathbf{Q}$ , the order statistic of  $\mathbf{S}$ , the labeling and the disease status corresponding to  $\tilde{\mathbf{Q}}$ , respectively. Moreover, let  $\tilde{Q}_k$  and  $\tilde{S}_k$ , denote the  $k$ th element  $\tilde{\mathbf{Q}}$  and  $\tilde{\mathbf{S}}$ , respectively.

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

## 2.2 Model

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

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

where  $Q_i = H(S_i)$ .

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

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

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

$$\text{AUC} = \Phi(a/\sqrt{1+b^2}) = \Phi(\mu/\sqrt{1+\sigma^2}). \quad (4)$$

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

$$P(L_i \neq 2 \mid Q_i, D_i) = g(Q_i), \quad (5)$$

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 (2005) proposed the following model:

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

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:

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

### 2.3 Prior distribution

We will follow a Bayesian approach to estimate  $(\lambda, \mu, \sigma)$ , equivalently,  $(\lambda, a, b)$ . 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  $(\mu, \sigma)$ . 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  $\mathbf{Q}$ , denoted by  $R(\mathbf{Q})$  and  $L(\mathbf{Q})$  respectively, are invariant after the transformation  $H$  on  $\mathbf{S}$ . Therefore, a set  $\mathcal{D}_{\text{obs}}$  invariant under the action of  $H$  can be defined as follows (Gu and Ghosal, 2009):

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

where

$$\begin{aligned} \mathbf{z} &= (z_1, \dots, z_m), \mathbf{w} = (w_1, \dots, w_n), \\ \underline{z}_k &= \max_i \{z_i : R_{Ni} < R_{Nk}\} \vee \max_j \{w_j : R_{N(m+j)} < R_{Nk}\}, \\ \bar{z}_k &= \min_i \{z_i : R_{Nk} < R_{Ni}\} \wedge \min_j \{w_j : R_{Nk} < R_{N(m+j)}\}, \\ \underline{w}_l &= \max_i \{z_i : R_{Ni} < R_{N(m+l)}\} \vee \max_j \{w_j : R_{N(m+j)} < R_{N(m+l)}\}, \\ \bar{w}_l &= \min_i \{z_i : R_{N(m+l)} < R_{Ni}\} \wedge \min_j \{w_j : R_{N(m+l)} < R_{N(m+j)}\}, \end{aligned}$$

for all  $k, i = 1, \dots, m; l, j = 1, \dots, n$ .

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.**

$$\int \bar{\Phi}(c_1 + c_2 t) \phi(t) dt = \bar{\Phi}(c_1 / \sqrt{1 + c_2^2}) \tag{9}$$

where  $c_2 \geq 0$ .

**Lemma 2.** Let  $S_1, \dots, S_N$  be the independent diagnostic variables with underlying disease

status  $D_1, \dots, D_N$  which follow Bernoulli ( $\lambda$ ). Assume that (2) and (5) hold. Then, we have

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

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

where  $L_i$  is defined in (1).

**Remark 1:** Lemma 2 implies that  $D$  and  $1\{L \neq 2\}$  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.

The following lemma will be used in proof of Theorem 1 in Section 3.

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

$$f_Q(t \mid L \neq 2) = (1 - \lambda_{(\mu,\sigma,g)}^*)\phi_{(g)}^*(t) + \lambda_{(\mu,\sigma,g)}^*\phi_{(\mu,\sigma,g)}^*(t), \quad (11)$$

where  $\lambda_{(\mu,\sigma,g)}^*$  and  $\phi_{(\mu,\sigma,g)}^*(t)$  are defined as

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

$$\phi_{(\mu,\sigma,g)}^*(t) = \frac{(1 - g(t))\phi_{(\mu,\sigma)}(t)}{\int (1 - g(s))\phi_{(\mu,\sigma)}(s)ds} \quad (13)$$

and  $\phi_{(g)}^*(t) = \phi_{(0,1,g)}^*(t)$ .

**Remark 2:** When (7) holds, (12) are (13) can be simplified to

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

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

**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

$$\begin{aligned} & \Pr\{(\mathbf{Z}, \mathbf{W}) \in \mathcal{D}_{\text{obs}} \mid \mu, \sigma\} \\ &= \Pr\{(\mathbf{z}, \mathbf{w}) \in \mathbb{R}^{m+n} : R(\mathbf{z}, \mathbf{w}) = R(\mathbf{S}), L(\mathbf{z}, \mathbf{w}) = L(\mathbf{S}) \mid \mu, \sigma\}. \end{aligned} \quad (16)$$

In the presence of the verification bias, the posterior distribution of  $(\lambda, \mu, \sigma)$ , 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  $\tilde{\mathbf{Q}}, \mathbf{D}_{\text{un}}$  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 stated as follows:

- Posterior distribution of  $(\mu, \sigma)$  given the rest:

$$\begin{aligned} \sigma^2 | \text{rest} &\sim \text{inverse gamma}((n' - 1)/2, (n' - 1)s_w^2/2), \\ \mu | \text{rest} &\sim \text{TN}(\bar{W}_{n'}, \sigma^2/n', (0, \infty)), \\ n' &= \sum_{i=1}^N \{1(\tilde{D}_i = 1, \tilde{L}_i = 2) + 1(\tilde{L}_i = 1)\}, \\ s_w^2 &= \sum_{j=1}^{n'} (W_j - \bar{W}_{n'})^2 / (n' - 1), \quad \bar{W}_{n'} = \sum_{j=1}^{n'} W_j / n'; \end{aligned} \tag{17}$$

- $\tilde{\mathbf{Q}}$  can be sequentially updated conditional on  $(\mu, \sigma)$  by

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

where  $i = 1, \dots, N$ ,  $\tilde{Q}_0 = -\infty$ ,  $\tilde{Q}_{N+1} = \infty$ .

- Posterior distribution of  $\lambda$  given the rest:

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

where  $m' = \sum_{i=1}^N \{1(\tilde{D}_i = 0, \tilde{L}_i = 2) + 1(\tilde{L}_i = 0)\}$ .

- Update the augmentation variable  $\tilde{D}_i \in \mathbf{D}_{\text{un}}$  by

$$\tilde{D}_i^{(\text{new})} \stackrel{\text{ind}}{\sim} \text{Ber} \left( \frac{\lambda \phi_{(\mu, \sigma)}(\tilde{Q}_i^{(\text{new})})}{\lambda \phi_{(\mu, \sigma)}(\tilde{Q}_i^{(\text{new})}) + (1 - \lambda) \phi(\tilde{Q}_i^{(\text{new})})} \right), \tag{20}$$

The posterior mean of  $(\mu, \sigma)$  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  $\pi^*(\tilde{\mathbf{Q}}, \mu, \sigma, \lambda, \mathbf{D}_{\text{un}} | \mathbf{S}, \mathbf{L})$ . Gibbs sampling procedure can be described as follows:

- (1) Choose an initial value of  $(\mu, \sigma)$ , say  $\mu = 1.5$  and  $\sigma = 1.5$ , and initialize  $\mathbf{D}_{\text{un}}$  by a realization of  $(N - N^*)$  subjects having  $\stackrel{\text{i.i.d.}}{\sim} \text{Bin}(p_0)$ , where  $p_0 = l_1 / (l_1 + l_0)$ . Conditional on the invariant set  $\mathcal{D}_{\text{obs}}$ , generate the initial values of  $Q_i \stackrel{\text{i.i.d.}}{\sim} \text{N}(0, 1)$  if  $D_i = 0$  and  $Q_i \stackrel{\text{i.i.d.}}{\sim} \text{N}(\mu, \sigma)$  if  $D_i = 1$ ,  $i = 1, \dots, N$ .
- (2) Start the iterations:
  - (a) Conditional on  $(\mu, \sigma)$ , update  $\tilde{\mathbf{Q}}$ , denoted as  $\tilde{\mathbf{Q}}^{(\text{new})}$ , with constraint  $(\mathbf{Z}, \mathbf{W}) \in \mathcal{D}_{\text{obs}}$  by following (18).
  - (b) Update  $\mathbf{Z}$  and  $\mathbf{W}$  values based on  $\tilde{\mathbf{Q}}^{(\text{new})}$  and  $\tilde{\mathbf{D}}$ .
  - (c) By following the posterior distributions specified in Section 2.4.3, update  $(\mu, \sigma, \lambda, \mathbf{D}_{\text{un}})$ .
- (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  $\mu/\sigma$  and  $1/\sigma$ , respectively. We can also calculate the  $100(1 - \alpha)\%$  credible interval for  $a$  as  $(q_{a, \alpha/2}, q_{a, 1-\alpha/2})$ , where  $q_{a, \alpha/2}$  and  $q_{a, 1-\alpha/2}$  denote  $\alpha/2$  and  $1 - \alpha/2$  quantiles of the sampled values of  $a$ . The credible interval of  $(q_{b, \alpha/2}, q_{b, 1-\alpha/2})$  for  $b$  is similarly defined. We also compute the estimate of AUC and its  $100(1 - \alpha)\%$  credible interval.

### 3. Consistency of the posterior

We prove posterior consistency only for the model (7). For other models, we can prove them in a similar way. For model (6), since there are no further parameters in the models, the proof is, in fact, simpler.

Let  $(\mu_0, \sigma_0, \alpha_0, \beta_0)$  be the true value of  $(\mu, \sigma, \alpha, \beta)$ . We will show the posterior for  $(\mu, \sigma, \alpha, \beta)$

is consistent at  $(\mu_0, \sigma_0, \alpha_0, \beta_0)$ , i.e., the posterior distribution concentrates around  $(\mu_0, \sigma_0, \alpha_0, \beta_0)$  (Ghosh and Ramamoorthi, 2003). Assume that the joint prior has a density  $\pi(\mu, \sigma, \alpha, \beta)$  with respect to the Lebesgue measure (denoted by  $\nu$ ).

**THEOREM 1:** *Assume (2) and (7) hold and  $\pi(\mu, \sigma, \alpha, \beta) > 0$  a.e.  $[\nu]$  over  $\mathbb{R} \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$ . Let  $(\mu_0, \sigma_0, \alpha_0, \beta_0)$  be the true value of  $(\mu, \sigma, \alpha, \beta)$ . Then for any neighborhood  $\mathcal{U}_0$  of  $(\mu_0, \sigma_0, \alpha_0, \beta_0)$ , we have that*

$$\lim_{N \rightarrow \infty} \pi^*((\mu, \sigma, \alpha, \beta) \in \mathcal{U}_0 | R(\mathbf{S}), \mathbf{L}) = 1 \quad a.s. \quad [P_{\mu_0, \sigma_0, \alpha_0, \beta_0, H}^\infty], \quad (21)$$

for  $(\mu_0, \sigma_0, \alpha_0, \beta_0)$  a.e.  $[\nu]$ , where  $P_{\mu_0, \sigma_0, \alpha_0, \beta_0, H}^\infty$  denotes the true joint distribution.

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  $(\mu, \sigma)$  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 of the Appendix). For a rank based method like BRL, this was shown by essentially major steps in Gu and Ghosal (2009). 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 (2009), 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 (A.1) 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 any unspecified parameters in the verification distribution (such as  $(\alpha, \beta)$  in (7)), which will need to be consistently estimated as well. This would be possible whenever the family of distributions of  $U_j$  corresponding to verified disease status satisfies Cramér-type regularity conditions with respect to the binormal parameters  $(\mu, \sigma)$  and all parameters in the verification distribution.

#### 4. Simulation studies

Lemma 2 implies that within a MCMC iteration, updating of  $\mathbf{D}_{\text{un}}$  will depend on  $(\lambda, \mu, \sigma, Q_i)$  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), full imputation (FI), mean score imputation (MSI), and inverse probability weighting (IPW) estimators proposed by Alonzo and Pepe (2005).

##### 4.1 Under bionormality assumption

Sets of simulations letting true AUC equal to 0.85 and 0.75, 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. The parameters  $\alpha = -0.53$  for AUC=0.75, and  $\alpha = -0.67$  for AUC=0.85, respectively, fixed  $\beta = 0.3$  in (7), and  $p_1 = 0.8$

and  $p_2 = 0.2$  in (6) are chosen 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. For each replication, there are total 100 and 200 subjects whose diagnostic test results are generated from  $N(0, 1)$  and  $N(\sqrt{1 + \sigma^2}\Phi^{-1}(\text{AUC}), \sigma^2)$  (see (4)) with a disease prevalence rate 0.25, where  $\sigma$  is equal to 1.5 for all cases, that is, for  $\text{AUC}=0.85$ , the true values of  $a$  and  $b$  are 1.2457 and 0.6667, respectively; for  $\text{AUC}=0.75$ , the true values of  $a$  and  $b$  are 0.8107 and 0.6667, respectively. No covariates are used. We compare our proposed PG-BRL method with SPE-E, PI, MSI and IPW methods in terms of accuracy.

From the simulation results (Table 1), we can see that PG-BRL performs consistently better in term of accuracy among these 5 methods. Especially when AUC decreases to 0.75, PG-BRL clearly shows more accuracy than the other methods. The corresponding PG-BRL estimates of  $(a, b)$  are shown in Table 2.

[Table 1 about here.]

[Table 2 about here.]

#### 4.2 Departure from binormality assumption

Since our method is based on binormality assumption, it is technically inapplicable if this assumption is violated. However, we have also studied the effect of departure from binormality. We generate  $(X_1, \dots, X_m)$  independently from beta distribution with mean 0.15 and standard deviation 0.15,  $(Y_1, \dots, Y_n)$  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 AUC estimate by PG-

BRL method has high bias than SPE-E and IPW methods, its MSE is much smaller. The simulation results show that our procedure has reasonable robustness properties against departure from binormality.

[Table 3 about 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 data base is still helpful. Here, the specified liver enzyme tests include serum total bilirubin (STB) level, 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 will 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. (2009), and Rockey et al. (2010), 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 4 and Figure 1.

[Table 4 about here.]

[Figure 1 about 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 cholestasis based on STB is more accurate than that based on AKP. That 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  $\text{TPR}(c)$  and  $\text{FPR}(c)$ ; see Figure 1 for the ROC curve estimates by PG-BRL, SPE-E, FI, MSI and IPW methods. The PG-BRL method does not suffer from this problem. For all the other approaches we used to compare with our method, it is difficult to know or estimate  $S_{(p_1N)}$  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 covariates  $\mathbf{V}$  are also observed, which carry a linear effect through an unknown regression effect,  $\boldsymbol{\beta}$ , i.e., the binormal model will become

$$\begin{aligned} H(X_i - \boldsymbol{\beta}^T \mathbf{V}_i) \mid D_i = 0 &\stackrel{\text{i.i.d.}}{\sim} N(0, 1), \\ H(Y_i - \boldsymbol{\beta}^T \mathbf{V}_i) \mid D_i = 1 &\stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2). \end{aligned}$$

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 covariates need 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 might 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 dependence of (5) on disease status will make expression in (10) dependent on the disease status as well as the functional form of  $g$  in (5).

## References

- Alonzo, T.A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society, Series C*, **54**, 173–190.
- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology*, **12**, 387–415.
- Begg, C. B. and Greenes, R. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, **39**, 207–215.
- Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B*, **60**, 71–87.

- Fluss, R. Reiser B., Faraggi D., and Rotnitzky A. (2009). Estimation of the ROC curve under verification bias. *Biometrical Journal*, **3**, 475–490.
- Fontana, R. J., Watkins, P. B., Bonkovsky, H. L., Chalasani, N., Davern, T., Serrano, J., Rochon, J. (2009). Drug-induced liver injury network (DILIN) prospective study: rationale, design and conduct. *Drug Saf*, **32**, 55–68.
- Ghosal, S. and Van der Vaart, A. W. (2011). Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press (to appear).
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). Bayesian Nonparametrics. New York: Springer-Verlag.
- Grant, A. and Neuberger, J. (1999). Guidelines on the use of liver biopsy in clinical practice. *British Society of Gastroenterology*, **Gut 45 Suppl 4**, IV1–CIV11.
- Gu, J. and Ghosal, S. (2009). Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference*, **139(6)**, 2076–2083.
- Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostics Imaging*, **29**, 307–335.
- Hájek, J. and Šidák, Z. (1967). Theory of Rank Tests. New York: Academic Press.
- Kosinski, A. S. and Barnhart, H. X. (2003). Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*, **59**, 163–171.
- Little R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. New York: John Wiley and Sons.
- Poynard, T., Imbert-Bismut, F., and Ratziu V. (2004). Serum markers of liver fibrosis. *Hepatology Reviews*, **1 (1)**, 23–31.
- Reilly, M. and Pepe, M. S. (1995). A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**, 299–314.
- Rockey, D. C., Seeff, L. B., Rochon, J. Freston, J., Chalasani, N., Bonacini, M., Fontana,

- R. J., Hayashi, P. H. (2010). Causality assessment in drug-induced liver injury using a structured expert opinion process: Comparison to the Roussel-Uclaf causality assessment method. *Hepatology*, **51**, 2117–2126.
- Rotnitzky, A., Faraggi, D., and Schisterman, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, **101**, 1276–1288.
- Zhou, X. H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communication in Statistics — Theory and Methods*, **22**, 3177–3198.
- Zhou, X. H. (1994). Effect of verification bias on positive and negative predictive values. *Statistics in Medicine*, **13**, 1737–1745.
- Zhou, X. H. (1998). Correcting for verification bias in studies of a diagnostic test’s accuracy. *Statistical Methods in Medical Research*, **7**, 337–353.

*Received June 2011. Revised February 2012.*

## Appendix

Proof of Lemma 1.

The result is well known, but we give a proof for completeness.

*Proof.* 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,

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

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

Proof of Lemma 2.

Since, we have  $f_{Q_i}(t | D_i = 1, L_i = 2)P(L_i = 2 | Q_i = t, D_i = 1) = \phi_{(\mu, \sigma)}(t)(1 - g(t))$  and  $f_{Q_i}(t | D_i = 0, L_i = 2)P(L_i = 2 | Q_i = t, D_i = 0) = \phi(t)(1 - g(t))$ , we have, by Bayes' theorem,

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

By following the same lines, we have

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

Proof of Lemma 3.

By Bayes' theorem, we have

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

where  $\lambda_{(\mu,\sigma,g)}^*$  and  $\phi_{(\mu,\sigma,g)}^*(t)$  are defined in (12) and (13), respectively.  $\square$

**THEOREM 2** (Doob's Theorem): [cf. Ghosal and Van der Vaart, 2011]

Let  $X^{(n)}$  be observations whose distribution depends on a parameter  $\theta$ , and both  $X^{(n)}$  and  $\theta$  take values in Polish spaces. Let  $\Pi$  be a prior distribution on  $\theta$ . Assume that  $\theta$  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  $\theta$  and  $(X^{(n)} : n \geq 1)$ . Then the posterior  $\Pi(\cdot \mid X^{(n)})$  is strongly consistent at  $\theta$  for almost every  $\theta$  [II].

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

Let  $1 \leq i_1 < \dots < i_{N^*} \leq N$  be the collection of indices for which  $L_{i_j} = 0$  or 1, and  $j = 1, \dots, N^*$ . From (14) and (15) in Remark 2, we know  $f_Q(t \mid L_i \neq 2) = (1 - \lambda^*)\phi_{\alpha,\beta}^*(t) + \lambda^*\phi_{\mu,\sigma,\alpha,\beta}^*(t)$ . Hence, unconditionally, we have  $Q_{i_j} \stackrel{\text{i.i.d.}}{\sim} (1 - \lambda^*)\Phi_{\alpha,\beta}^* + \lambda^*\Phi_{\mu,\sigma,\alpha,\beta}^*$ , where  $\Phi_{\alpha,\beta}^*(t)$ , and  $\Phi_{(\mu,\sigma,\alpha,\beta)}^*$  are the CDFs of  $\phi_{\alpha,\beta}^*(t)$  and  $\phi_{(\mu,\sigma,\alpha,\beta)}^*(t)$  given by (15), respectively. Thus we

have

$$U_j = ((1 - \lambda^*)\Phi_{\alpha,\beta}^* + \lambda^*\Phi_{(\mu,\sigma,\alpha,\beta)}^*)(Q_{i_j}) \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1). \quad (\text{A.1})$$

Let  $(R'_{N^*1}, \dots, R'_{N^*N^*})$  be the rank vector of  $(U_1, \dots, U_{N^*})$ , and  $(L'_{N^*1}, \dots, L'_{N^*N^*})$  stand for their labels. Now, as in Theorem *a* on page 157 of Hájek and Šidák (1967), we have

$$\begin{aligned} \mathbb{E}\left(U_j - \frac{R'_{N^*i_j}}{N^* + 1}\right)^2 &= \frac{1}{N^*} \sum_{k=1}^{N^*} \mathbb{E}\left[\left(U_j - \frac{k}{N^* + 1}\right)^2 \middle| R'_{N^*i_j} = k\right] \\ &= \frac{1}{N^*} \sum_{k=1}^{N^*} \frac{k(N^* - k + 1)}{(N^* + 1)^2(N^* + 2)} < \frac{1}{N^*} \rightarrow 0 \text{ as } N \rightarrow \infty, \end{aligned}$$

where the expectation is interpreted as conditional on the labels. Therefore,

$$U_j = \lim_{k \rightarrow \infty} \frac{R'_{N_k^*i_j}}{N_k^* + 1} \quad (\text{A.2})$$

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

Let  $(Q_{i_1}^*, \dots, Q_{i_{m^*}}^*) = (Q_{i_j}^* : L_{i_j} = 0, j = 1, \dots, m^*)$ . Hence for  $j = 1, \dots, m^*$ ,  $U_j^* = (1 - \lambda^*)\Phi_{\alpha,\beta}^* + \lambda^*\Phi_{(\mu,\sigma,\alpha,\beta)}^*(Q_{i_j}^*) \stackrel{\text{i.i.d.}}{\sim} g_{\mu,\sigma,\alpha,\beta}$  (say), which belongs to a regular parametric family indexed by parameters  $(\mu, \sigma, \alpha, \beta)$ . More specifically, Cramér-type regularity conditions can be verified by applying the inverse function theorem. Thus some consistent estimator for  $(\mu, \sigma, \alpha, \beta)$ , such as the MLE, a Bayes estimator, the method of moment estimator, or a minimum distance estimator, can be easily obtained. Hence, there exists a function  $h : [0, 1]^\infty \rightarrow \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$ , such that  $(\mu, \sigma, \alpha, \beta) = h(U_1^*, U_2^*, \dots)$ .

Therefore, there exists a function  $h^*$  of all ranks and observed labels such that

$$\begin{aligned} (\mu, \sigma, \alpha, \beta) &= h(U_1^*, U_2^*, \dots) \\ &= h(h_1(\mathbf{R}_N, N \geq 1, L_1, L_2, \dots), h_2(\mathbf{R}_N, N \geq 1, L_1, L_2, \dots), \dots) \\ &= h^*(\mathbf{R}_N, N \geq 1, L_1, L_2, \dots) \quad \text{a.s. } [P_{\mu_0, \sigma_0, \alpha_0, \beta_0, H}^\infty]. \end{aligned}$$

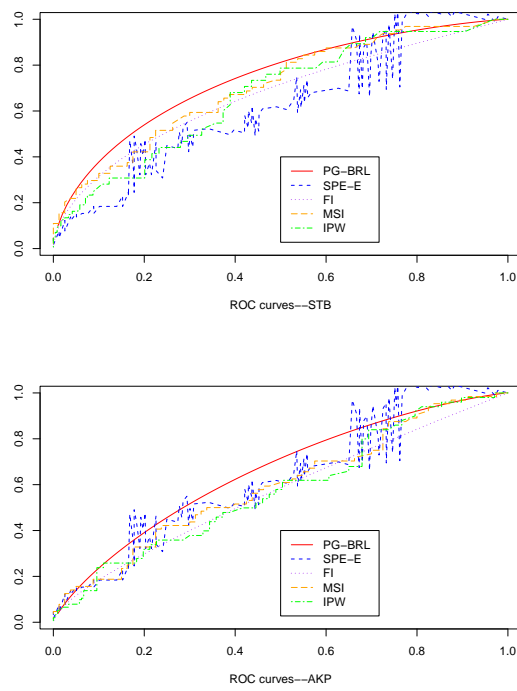


Figure 1. Real data analysis: ROC curve estimates for STB and AKP.

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

Imputation		$N = 100$		$N = 200$	
Model	Method	Bias	MSE	Bias	MSE
Probit AUC=.85	PG-BRL	-.0274	.0088	-.0115	.0041
	SPE-E	.0190	.0075	.0157	.0035
	FI	.0391	.0054	.0359	.0032
	MSI	.0346	.0055	.0316	.0031
	IPW	-.0309	.0110	-.0330	.0056
Probit AUC=.75	PG-BRL	.0062	.0062	.0077	.0042
	SPE-E	.0243	.0116	.0216	.0056
	FI	.0566	.0107	.0526	.0061
	MSI	.0495	.0103	.0458	.0057
	IPW	-.0291	.0137	-.0306	.0068
Threshold AUC=.85	PG-BRL	-.0007	.0074	-.0014	.0042
	SPE-E	.0205	.0096	.0056	.0053
	FI	.0429	.0059	.0344	.0036
	MSI	.0371	.0058	.0291	.0035
	IPW	.0121	.0097	.0021	.0054
Threshold AUC=.75	PG-BRL	.0234	.0076	.0158	.0049
	SPE-E	.0230	.0150	.0117	.0072
	FI	.0627	.0112	.0572	.0069
	MSI	.0543	.0108	.0493	.0063
	IPW	.0117	.0155	.0064	.0074

**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 AUC=.85	PG-BRL $a$	.0641	.5383	.0390	.2276
	$b$	.0120	.0959	.0049	.0447
Probit AUC=.75	PG-BRL $a$	.1146	.2579	.0817	.1253
	$b$	-.0094	.0682	.0096	.0329
Threshold AUC=.85	PG-BRL $a$	.2919	.7202	.1415	.3405
	$b$	.1128	.1172	.0527	.0541
Threshold AUC=.75	PG-BRL $a$	.2385	.4177	.1327	.1755
	$b$	.0551	.0810	.0334	.0352

**Table 3**

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

Imputation		$N = 200$	
Model	Method	Bias	MSE
Probit AUC=.715	PG-BRL	0.0257	.0022
	SPE-E	.0070	.0040
	FI	-.056	.0076
	MSI	-.0381	.0055
	IPW	.0066	.0042
Threshold AUC=.715	PG-BRL	0.0268	.0026
	SPE-E	.0063	.0057
	FI	-.0930	.0125
	MSI	-.0746	.0093
	IPW	.0041	.0055

**Table 4**

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

Lab Test	Method	$\widehat{AUC}(sd)$	95% CI
STB	PG-BRL	0.7388(0.0559)	(0.6389,0.8601)
	SPE-E	0.7198(0.0655)	(0.597,0.8495)
	FI	0.6704(0.0357)	(0.6055, 0.7337)
	MSI	0.7041(0.0417)	(0.6201,0.7849)
	IPW	0.6714(0.0481)	(0.579,0.7643)
AKP	PG-BRL	0.6550(0.0738)	(0.5624,0.8527)
	SPE-E	0.6283(0.0962)	(0.4421,0.8183)
	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)