ARTICLE

Bayesian nonparametric estimation of ROC surface under verification bias

Rui Zhu  |  Subhashis Ghosal

1Department of Statistics, North Carolina State University, North Carolina, USA

Correspondence
Rui Zhu, Department of Statistics, North Carolina State University. Email: rzhu4@ncsu.edu

Present Address
Department of Statistics, North Carolina State University

Summary
The Receiver Operating Characteristic (ROC) curve has been widely used to assess the accuracy of a diagnostic test for three categories. A common problem is verification bias, referring to the situation where not all subjects have their true classes verified. In this paper, we consider the problem of estimating the ROC surface under verification bias. We adopt a Bayesian nonparametric approach by directly modeling the underlying distributions of the three categories by Dirichlet Process mixture priors. We propose a robust computing algorithm by only imposing a missing at random assumption for the verification process but no assumption on the distributions. The method can also accommodate covariates information in estimating the ROC surface, which can lead to a more comprehensive understanding of the diagnostic accuracy. It can be adapted and hugely simplified in the case where there is no verification bias, and very fast computation is possible through the Bayesian bootstrap process. The proposed method is compared with other commonly used methods by extensive simulations. We find that the proposed method generally outperforms other approaches. Applying the method to two real datasets, the key findings are as follows: (i) HE4 has a slightly better diagnosis ability compared to CA125 in discriminating healthy, early stage and late stage patients of epithelial ovarian cancer. (ii) Serum albumin has a prognostic ability in distinguishing different stages of hepatocellular carcinoma.

KEYWORDS: ROC surface, Verification bias correction, Dirichlet Process, Bayesian bootstrap, MAR assumption

1 INTRODUCTION

The Receiver Operating Characteristic (ROC) curve has been widely accepted as a useful tool in assessing the accuracy of a diagnostic test and it has been studied since the 1950s. By putting the True Positive Rate on the y-axis and the False Positive Rate on the x-axis, this curve shows how those rates change when the cut-off point for decision varies. The Area Under the ROC Curve (AUC) is considered to be the most commonly used index to quantify the performance of a diagnostic test.

However, one limitation of the ROC curve is that it can only be applied to binary classifiers. In reality we may encounter the cases where diagnostic decisions should be applied to multiple classes. For example, Alzheimer’s disease has an early stage which is represented by a mild cognitive impairment. It is the transition stage between normal aging and Alzheimer’s disease[1]. At this stage, the patients may experience memory loss and difficulty in problem solving. The symptoms for this stage is clearly
different from late-stage Alzheimer's disease, and so the treatment should be different as well. Because of that, we need a biomarker to distinguish three stages: healthy, early stage and late stage. The ROC curve methodology and the related properties are no longer applicable in this case. A natural generalization of the ROC curve to three classes, termed as the ROC surface, plots the True Class Fractions (TCFs) in three axes respectively, and thus illustrates the trade-off among the three TCFs as the cut-off points vary. More specifically, let $X$, $Y$, $Z$ denote the measurements from three classes respectively, and $F_0$, $F_1$, $F_2$ denote their underlying distribution functions. Given the two cut points $c_1$ and $c_2$, with $c_1 < c_2$, the TCFs will be defined as follows:

$$
\text{TCF}_1 = P(X \leq c_1) = F_0(c_1), \quad \text{TCF}_2 = P(c_1 < Y \leq c_2) = F_1(c_2) - F_1(c_1), \quad \text{TCF}_3 = P(Z > c_2) = 1 - F_2(c_2)
$$

(1)

The ROC surface is constructed by plotting the co-ordinates of (TCF$_1$, TCF$_2$, TCF$_3$). The functional representation of the ROC surface is given as follows:

$$
\text{ROC}_s(p_1, p_3) = \left\{ \begin{array}{ll}
F_1(F_2^{-1}(1 - p_3)) - F_1(F_0^{-1}(p_1)), & \text{if } F_0^{-1}(p_1) \leq F_2^{-1}(1 - p_3) \\
0, & \text{otherwise}
\end{array} \right.
$$

(2)

Analogous to the AUC, the volume under the surface (VUS) is proposed as a measure of the accuracy for three-dimensional classifiers, which can be calculated by integrating out the ROC surface, i.e.,

$$
\text{VUS} = \int_0^1 \int_0^1 \int_0^1 \text{ROC}_s(p_1, p_3) dp_3 dp_1.
$$

The VUS takes values in $[1/6, 1]$. The larger the VUS is, the better discriminatory ability the test has. For a perfect test, that is, a test that can correctly classify every instance, $\text{VUS} = 1$; for a test no better than random guessing, $\text{VUS} = 1/6$.

There are several approaches to estimating the ROC surface as well as the VUS. The empirical estimator of the ROC surface was obtained by simply replacing the cumulative distribution functions in (2) by their empirical estimators and the VUS can be calculated by integrating out this ROC estimator. Another empirical estimator was given by the unbiased nonparametric Mann-Whitney U-statistic of the probability $P(X < Y < Z)$ which was later extended to the case with ties. A kernel estimator was proposed to get a smooth estimate of the surface using the Gaussian kernel. A Bayesian nonparametric estimator has been proposed by Inácio et al. (2011) based on the mixtures of finite Pólya trees priors. Another Bayesian nonparametric method using the Bayesian bootstrap technique has been proposed by de Carvalho et al. (2018) who extend the method proposed by Gu et al. (2008) from two dimensions to three.

Parametric methods and semi-parametric methods were also considered for estimating the ROC surface and the VUS. Xiong et al. (2006) incorporated the most common and widely used parametric trinormal assumption, i.e., $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$, and $Z \sim N(\mu_3, \sigma_3)$, where $N(\mu, \sigma)$ stands for a normal distribution with mean $\mu$ and standard deviation $\sigma$. Li and Zhou (2009) proposed a semi-parametric estimation using the ROC surface under the trinormal model. Zhu and Ghosal (2018) proposed a Bayesian semi-parametric estimation under the assumption that the underlying distributions are trinormal under some unknown transformation (called the trinormality assumption).

Another common problem in the data that complicates the ROC surface estimation is called verification bias. In reality, the true class each individual belongs to may not be completely known to us. In biomedical settings, the class may be the true disease status of the patient, which is usually verified through a gold standard test, the perfect existing diagnostic test. However, such a test may be expensive and invasive. It is more ethical to apply a gold standard test only to those high risk subjects according to the screening test. This leads to the problem of missing data. Because patients who get lower scores in screening test are more likely to have their labels missing, simply ignoring this missingness and estimating the ROC surface using only the labeled subjects may lead to biased results. One reasonable assumption is to treat the labels as Missing At Random (MAR). According to this assumption, the probability of verification (that is, observing the true label) does not depend on the disease status given the observed measurements. Because the true disease status is unknown to doctors when making the decision whether to refer patients for the gold standard test or not, this assumption should be reasonable in this context.

The number of papers in the literature addressing verification bias for the ROC surface is very limited. Chi and Zhou (2008) proposed using the maximum likelihood estimator to estimate the ROC surface and the VUS for ordinal responses under verification bias. To Duc et al. (2016) proposed several bias-corrected estimators of TCFs and then constructed bias-corrected ROC surfaces for continuous diagnostic tests. These methods were extended from Full Imputation (FI), Mean Score Imputation (MSI), Inverse Probability Weighted (IPW), Semi-Parametric Efficient (SPE) estimators for estimating the ROC curves proposed by Alonzo and Pepe (2005). Parametric models should be chosen to compute the probability for belonging to each class and also the probability of verification to adjust for the influence caused by missing labels. These methods addressed the problem of
estimating the ROC surface for continuous measurements under verification bias, but cannot lead to a smooth estimate which is often desired. Zhu and Ghosal (2018) also developed a Bayesian semi-parametric estimation method for verification bias. Through this method, they obtained a smooth surface. However, because of the trinormality assumption on which this method is based (please refer to the Supplementary Materials for more detail about the trinormality assumption), if the data is too far away from this assumption, this method may fail to give a reasonable result, which will be shown in the simulation section.

In this paper, we propose a new nonparametric Bayesian method for estimating the ROC surface and the VUS for continuous measurements by using the Dirichlet process to construct a prior for the distribution functions. We can produce a smooth surface and we do not impose any assumption on the data distributions. Moreover, our method can easily accommodate covariates in estimating the ROC surface. Covariates information like gender, age, height, weight, ethnicity is usually available under biomedical settings since the patients need to fill out the information when doing a prognostic test. It is absolutely a good idea to include them for assessing the accuracy of a diagnostic test since the distributions for three classes can be different conditional on the covariates. None of the methods mentioned above have considered incorporating the covariates in estimating the ROC surface in the presence of verification bias. To Duc et al. (2016) used covariates for building parametric models, but not directly on building ROC surface. Our method, on the other hand, can directly accommodate covariates by expressing ROC surface as a function of the covariates, and then get the covariate-adjusted ROC surface by integrating out the covariates. This can lead to a more comprehensive understanding of the diagnostic accuracy.

The rest of the paper is organized as follows: we start with the method under verification bias, and then extend this method to adjust covariates. Afterward we simplify this method to a computationally easier version for the case without verification bias. We then run extensive simulation studies in Section 5 and apply this method to two real datasets in Section 6. We first apply the method to compare the two biomarkers—CA125 and HE4 on their performances in discriminating healthy subjects, the early stage patients and the late stage patients of the epithelial ovarian cancer. We then apply this method to assess the ability of serum albumin in distinguishing different stages of hepatocellular carcinoma when incorporating gender as a covariate.

2 | METHOD UNDER VERIFICATION BIAS

2.1 | Notation

Denote the diagnostic measurements from the whole population by \( S = (S_1, \ldots, S_N) \), where \( N \) is the total number of subjects involved in the study. Among \( N \) observations, we have \( n_0 \) observations from the healthy group, \( n_1 \) observations from the level-1 disease group and \( n_2 \) observations from the level-2 disease group, \( N = n_0 + n_1 + n_2 \). Note that \( n_0, n_1, n_2 \) are unknown under verification bias. The true disease status of those \( N \) individuals are denoted by \( D = (D_1, \ldots, D_N) \). Only a small proportion of the true disease statuses are verified through the gold standard test. Let \( L = (L_1, \ldots, L_N) \) carry information on missingness for all subjects, as well as the true disease status if observed. Define a label variable \( L_i \) as follows:

\[
L_i = \begin{cases} 
  d, & \text{if label is observed and } D_i = d, \quad d = 0, 1, 2, \\
  3, & \text{if label is not observed.} 
\end{cases}
\]  

(3)

2.2 | Model

We assume the disease prevalence rates for level-1 and level-2 disease in the population are \( \lambda_1 \) and \( \lambda_2 \) respectively, where \( 0 < \lambda_1 < 1 \) and \( 0 < \lambda_2 < 1 \), \( \lambda_0 = 1 - \lambda_1 - \lambda_2 \), so \( (n_0, n_1, n_2) \sim \text{Mult}(N, (\lambda_0, \lambda_1, \lambda_2)) \). Conditional on the true disease labels, we have

\[
S_i | \mu_i, \sigma_i^2 \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad (\mu_i, \sigma_i^2) | \{D_i = d\} \sim G_d, \quad d = 0, 1, 2, 
\]

(4)

where \( \mathcal{N}(\mu_i, \sigma_i^2) \) stands for the normal distribution with mean \( \mu_i \) and variance \( \sigma_i^2 \).

Based on (4), the distribution functions of the observations for the healthy group, level-1 disease group and level-2 disease group (denoted by \( F_0, F_1, F_2 \), are mixtures of normal distributions with mixing distributions \( G_0, G_1 \) and \( G_2 \) respectively, and the density functions \( f_0, f_1, f_2 \) are mixtures of normal densities, i.e.,

\[
F_d = \int \Phi\left(\frac{x - \mu}{\sigma}\right) dG_d(\mu, \sigma), \quad f_d(x) = \int \phi\left(\frac{x - \mu}{\sigma}\right) dG_d(\mu, \sigma), \quad d = 0, 1, 2,
\]

(5)

where \( \Phi(\cdot) \) and \( \phi(\cdot) \) stand for the distribution function and density function respectively of a standard normal random variable.
We can express the points in the ROC surface as \((F_0(c_1), F_1(c_2) - F_1(c_1), 1 - F_2(c_2))\), where \(c_1 \in \mathbb{R}\) and \(c_2 \in \mathbb{R}\) are the two cut points for classifying the three categories. By varying the values of two cut points, we can obtain the ROC surface.

We also need to build a model for the verification process. We follow Gu et al. (2014) by making a very general assumption to model the true disease status as missing at random, i.e.,

\[
P(L_i \neq 3|S_i, D_i) = g(S_i), \quad i = 1, \ldots, N,
\]

where \(g\) is an increasing function. This is reasonable since, in clinical practice, the true disease status is verified through the gold standard test, which will typically be prescribed to patients according to the screening test results. That is to say, whether patients have their true disease status verified or not only depends on the observed diagnostic test but can depend on the true disease status only through its influence on the diagnostic test. A patient with a higher score in the observed diagnostic test has a higher chance to have the disease and thus is more likely to be forwarded to the more accurate test.

### 2.3 Prior distribution

To design a Bayesian algorithm, first of all we need to specify the prior distributions for distributions \(G_0, G_1\) and \(G_2\). We consider Dirichlet process priors (See Definition 4.1 in the Supplementary Materials) on these three distributions. Let \(\text{DP}(M F)\) stand for the Dirichlet process with prior precision \(M > 0\) and centering measure \(F\), we have:

\[
G_d \sim \text{DP}(M_d G_d^*), \quad G_d^* : \text{NIG}(m_d, a_d, s_d, \beta_d) \quad d = 0, 1, 2,
\]

where \(\text{NIG}(m_d, a_d, s_d, \beta_d)\) stands for the normal-inverse gamma distribution, i.e., \(\sigma^{-2} \sim \text{Ga}(s_d, \beta_d), \mu|\sigma \sim N(m_d, \sigma^2/a_d)\), \(M_d, s_d, \beta_d, a_d, d \in \{0, 1, 2\}\) are treated as constants which may be determined from the prior knowledge if available. The prior precision parameter \(M_d\) controls the variability of \(G_d\) around \(G_d^*\), with larger values of \(M_d\) leading to realizations of \(G_d\) that are closer to \(G_d^*\).

### 2.4 Posterior distribution

We apply Gibbs sampling technique to sample from the posterior distributions. Let \(\theta_i\) denote \((\mu_i, \sigma_i^2)^T\). Because \(\theta_i\) is closely related to \(D_i\), we choose to update them as a block. We use subscript \(-i\) to denote the set of all index \(j \neq i\). For example, \(\theta_{-i} = (\theta_j : j \neq i), D_{-i} = (D_j : j \neq i)\). The posterior distributions can be described as follows:

- The posterior distribution of \(D_i\) given \(S, \theta_{-i}, D_{-i}, L_i = 3\) is given according to the Bayes rule:

\[
\left(\mathbb{I}\{D_i = 0\}, \mathbb{I}\{D_i = 1\}, \mathbb{I}\{D_i = 2\}\right) | (S, \theta_{-i}, D_{-i}, L_i = 3, G_0, G_1, G_2) \sim \text{Mult}\left(1, \left(\frac{\lambda_0 f_0(S_i)}{f(S_i)}, \frac{\lambda_1 f_1(S_i)}{f(S_i)}, \frac{\lambda_2 f_2(S_i)}{f(S_i)}\right)\right),
\]

where \(\lambda_0, \lambda_1, \lambda_2\) are the disease prevalence rates for healthy, level-1 and level-2 disease in the population, \(f_0, f_1, f_2\) are defined in (5) above and \(f(t) = \lambda_0 f_0(t) + \lambda_1 f_1(t) + \lambda_2 f_2(t)\). However, notice that \(G_0, G_1\) and \(G_2\) are unknown. We can only generate a set of samples of \(G_0, G_1\) and \(G_2\) from their posterior distributions given \(\theta_{-i}, D_{-i}\) and then plug in (5) to calculate \(f_0, f_1\) and \(f_2\). Take \(G_0\) as an example. The posterior distribution of \(G_0\) given \(\theta_{-i}, D_{-i}\) is \(\text{DP}(M_0 G_0^* + \sum_{j \neq i} \delta_{\theta_j} \mathbb{I}\{D_j = 0\})\) according to Theorem 4.6 in Ghosal and van der Vaart (2017) (see Supplementary Materials). The sampling process for this posterior distribution is as follows:

1. According to Proposition G.10 of Ghosal and van der Vaart (2017) (see Supplementary Materials), generate \(V \sim \text{Be}(M_0, n_0^*)\) first, where \(n_0^* = \sum_{j \neq i} \mathbb{I}\{D_j = 0\}\).
2. With probability \(V\), we generate a sample from \(\text{DP}(M_0 G_0^*)\) using Theorem 4.19 (ii) in Ghosal and van der Vaart (2017) (see Supplementary Materials). An appropriate sample is obtained by calculating \(\sum_{j \neq i} q_j \delta_{\theta_j}\), where \(N^* \gg M_0\) is a constant, \((q_1, \ldots, q_{N^*}) \sim \text{Dir}(N^*, M_0/N^*, \ldots, M_0/N^*)\), and \(\theta_j \overset{i.i.d.}{\sim} G_0^*\) for \(j = 1, \ldots, N^*\).
3. With probability \(1 - V\), generate a sample from \(\sum_{j \neq i} \delta_{\theta_j}\), i.e., calculate \(\sum_{k=1}^{n_0^*} w_k \delta_{\theta_k}\), where \((w_1, \ldots, w_{n_0^*}) \sim \text{Dir}(n_0^*; 1, 1, \ldots, 1)\), \(j_k\) satisfies that \(D_{j_k} = 0, k = 1, \ldots, n_0^*\).

Plugging this sample in (5), we get \(f_0\). Similarly for \(f_1\) and \(f_2\).
The three distribution functions of memory needed, the output will be stored by the function values on the grid points, i.e., the outputs are the mean value of \( m_i \) for each disease prevalence rate. In the simulation part in Section 5, we will use a Bayes classifier to determine the initial value for \( D_i \) when \( L_i \neq 3 \). We monitor the trace plot to make sure the algorithm converges and discard all samples before a suitable burn-in period.

Remark 1. Here we assume that the disease prevalence rates \( (\lambda_0, \lambda_1, \lambda_2) \) are known. The disease prevalence rates in the population can be found in historical data. If the data we are looking at is representative of the population, using the population level disease prevalence rates is a good choice. We can also treat \( (\lambda_0, \lambda_1, \lambda_2) \) as unknown and assign some priors on them. The likelihood is given by

\[
P(S, D | \lambda_0, \lambda_1, \lambda_2) = \prod_{i = D_i = 0} f_0(S_i) \prod_{i : D_i = 1} f_1(S_i) \prod_{i : D_i = 2} f_2(S_i) \prod_{i : D_i = 3} \left\{ \hat{\lambda}_0 f_0(S_i) + \lambda_1 f_1(S_i) + \lambda_2 f_2(S_i) \right\}
\]

We can update the posterior distributions accordingly.

A computational algorithm can be developed following the posterior distributions given above. The algorithm is given in the Supplementary Materials. Notice that when \( L_i \neq 3 \) then \( D_i = L_i, i = 1, \ldots, N \), we need to assign an initial value for \( D_i \) when \( L_i = 3 \). We use a Bayes classifier to determine the initial value for \( D_i \) when \( L_i = 3 \), i.e., \( D_i = \arg\max_{j=0,1,2} P(S_i | D_i = j) \), where \( P(S_i | D_i = j) \) is a rough estimate given by a kernel estimator based on the verified subjects.

The hyperparameters \( \mathbf{m}, \mathbf{a}, \mathbf{s}, \mathbf{b} \) are chosen according to our prior knowledge about the distributions for different classes. If the prior information is not available, we will choose them to be non-informative. In the simulation part in Section 5, \( m_d \) is chosen according to the mean value of \( S_i \) whose \( L_i = d \), where \( d = 0, 1, 2 \). For others, since the prior information is not available, the parameters \( \mathbf{M} \) and \( \mathbf{a} \) are chosen to be \( (1, 1, 1) \), \( \mathbf{s} \) and \( \mathbf{b} \) are chosen to be \( (0.1, 0.1, 0.1) \), \( N^* \) is chosen to be 1000. To reduce the computation memory needed, the output will be stored by the function values on the grid points, i.e., the outputs are \( \hat{F}_0, \hat{F}_1, \hat{F}_2 \), where \( \hat{F}_d = (F_{d,1}, \ldots, F_{d,niter}) \) for \( d = 0, 1, 2 \). The distribution functions are stored on a grid of points \( (C_1, C_2, \ldots, C_K) \), i.e., record the values of \( (F_{0,m}(C_1), \ldots, F_{0,m}(C_K), F_{1,m}(C_1), \ldots, F_{1,m}(C_K), F_{2,m}(C_1), \ldots, F_{2,m}(C_K)) \) for each iteration, \( m = 1, 2, \ldots, niter \).

We monitor the trace plot to make sure the algorithm converges and discard all samples before a suitable burn-in period. The three distribution functions \( \hat{F}_0, \hat{F}_1, \hat{F}_2 \) are finally estimated according to the sample means of the values in each grid point.

---

1. The code for this method along with the real data analysis can be found in https://github.com/RrZzZz/Bayesian-nonparametric-estimation-of-ROC-surface-under-verification-bias
(C_1, C_2, \ldots, C_K). That is, \( F_0(C_1), \ldots, F_0(C_K) \) for \( F_0, F_1(C_1), \ldots, F_1(C_K) \) for \( F_1 \) and \( F_2(C_1), \ldots, F_2(C_K) \) for \( F_2 \). The ROC surface is estimated based on a \( K \times K \) grid points given by \( (F_0(c_1), F_1(c_2) - F_1(c_1), 1 - F_2(c_2)) \), where \( c_1 \) and \( c_2 \) can take values in \( (C_1, C_2, \ldots, C_K) \). The VUS can then be calculated based on this ROC surface.

### 3 | METHOD UNDER VERIFICATION BIAS WITH COVARIATES

Generally in biomedical settings, apart from the test results, we also have some covariates information. Typical covariates can be patients’ gender, age, height, weight, ethnicity etc. The accuracy of the diagnostic test can vary for different covariates values. So we also want to incorporate this information when estimating the ROC surface. The methodology is very similar to the one without covariates, just a few adjustments to the model are needed.

Denote the covariates for patient \( i \) by \( w_i = (1, w_{i1}, \ldots, w_{ik})^T \), where \( k \) is the number of covariates we have. Assume that the observed diagnostic measurement \( S_i \) has a linear relationship with the covariates \( w_i \), i.e.,

\[
S_i | \gamma_i, w_i, \sigma^2_j \sim N(\gamma_i^T w_i, \sigma^2_j), \quad (\gamma_i, \sigma^2_j) | \{D_i = d\} \sim G_d, \quad d = 0, 1, 2,
\]

(12)

Notice that the parameters \( \gamma_i \)s and \( \sigma^2_j \)s are specific to each individual \( i \). The distribution functions and density functions of the observations for each group will depend on the covariates \( w_i \). Let \( f_{0w} \) denote the conditional distribution function for observations in the healthy group conditional on the covariates, \( f_{1w} \) denote the conditional distribution in level-1 disease group and \( f_{2w} \) denote the conditional distribution in level-2 disease group. The conditional density functions are \( f_{0w}, f_{1w} \) and \( f_{2w} \) respectively. Based on (12), we have:

\[
F_{dw}(x) = \int \Phi(\frac{x - \gamma^T w}{\sigma})dG_d(\gamma, \sigma), \quad f_{dw}(x) = \int \phi(\frac{x - \gamma^T w}{\sigma})dG_d(\gamma, \sigma), \quad d = 0, 1, 2,
\]

which gives a semiparametric model.

Following Li et al. (2012), the covariate-specific ROC surface is defined as

\[
\text{ROC}_{w}(p_1, p_2) = \begin{cases} F_{1w}(F_{2w}^{-1}(1 - p_3)) - F_{1w}(F_{0w}^{-1}(p_1)), & \text{if } F_{0w}^{-1}(p_1) \leq F_{2w}^{-1}(1 - p_3) \\ 0, & \text{otherwise} \end{cases}
\]

(14)

and the covariate-adjusted ROC surface is defined as

\[
\text{ROC}(p_1, p_3) = \int \text{ROC}_{w}(p_1, p_3)dJ_w(w)
\]

(15)

where \( J_w(w) \) stands for the distribution function for \( w \).

The prior distributions are adjusted to be

\[
G_d \sim \text{DP}(M_d G_d^*), \quad G_d^* : \text{NIG}(\gamma_{0d}, \Lambda_d, s_d, \beta_d) \quad d = 0, 1, 2,
\]

where \( \text{NIG}(m_d, a_d, s_d, \beta_d) \) still stands for the normal inverse gamma distribution, i.e., \( \sigma^{-2} \sim \text{Ga}(s_d, \beta_d), \gamma | \sigma \sim N(\gamma_{0d}, \sigma^2 \Lambda_d^{-1}) \), here \( \gamma_{0d} \) is a \((k + 1) \times 1\) vector and \( \Lambda_d \) is a \((k + 1) \times (k + 1)\) matrix for \( d = 0, 1, 2 \).

Let \( \theta_i = (\gamma_i, \sigma_i^2) \), the posterior distribution of \( \theta_i \) conditional on \( D, L, S, \theta_{-i} \) is still given by

\[
\theta_i | (\theta_{-i}, S, D_i = d) \sim \sum_{j \neq i} q_{i,j} \delta_{\theta_j} \{D_j = d\} + q_{i,d} G_{b_i},
\]

(17)

where \( d \in \{0, 1, 2\} \), the probability vector \( (q_{i,j} : j \in \{1, \ldots, N\} \setminus \{i\}) \) is now satisfying

\[
q_{i,j} \propto \begin{cases} \frac{M_d \beta_{0d}^+ \Gamma(s_d + 1/2)\sqrt{|\Lambda_d|}}{(2\pi)^{1/2} \Gamma(s_d)} \left[ \beta_d + \frac{1}{2} \left( S_i^2 + \gamma_{0d}^T \Lambda_d s_d \gamma_{0d} - \gamma_{0d}^T \Lambda_d \gamma_{0d}^* \right) \right]^{-(s_d + 1/2)}, & j = 0, \\ \sigma_j^{-1} \exp \left\{ -(S_i - \gamma_j^T w_i)^2 / (2\sigma_j^2) \right\}, & \text{otherwise} \end{cases}
\]

\[D_j = d, j \neq i, j \geq 1, \text{ otherwise,}\]

(18)

and \( G_{b_i} \) is given by

\[
\gamma | \sigma, S_i \sim N(\gamma_{0d}, \sigma^2 \Lambda_d^{-1}),
\]

\[
\sigma^{-2} | S_i \sim \text{Ga}(s_d + 1/2, \beta_d) + \frac{1}{2} \left( S_i^2 + \gamma_{0d}^T \Lambda_d s_d \gamma_{0d} - \gamma_{0d}^T \Lambda_d \gamma_{0d}^* \right),
\]

(19)

where \( \gamma_{0d}^* = (\Lambda_d + w_i w_i^T)^{-1}(\Lambda_d s_d + S_i w_i) \), \( \Lambda_d^* = \Lambda_d + w_i w_i^T \).
4 | METHOD WITHOUT VERIFICATION BIAS

The above Bayesian nonparametric method can be substantially simplified for the case without verification bias by skipping the step to update \( D_i \) since, under the gold standard, all \( D_i \)'s are known to us. As no classification step will be needed, the kernel smoothing step will also be unnecessary, and hence the Dirichlet process can be directly applied on the distributions of \( X, Y \) and \( Z \). However, then the sampling process will be very close to the Bayesian bootstrap process, which is much easier to sample from with only a fixed number of independent exponential random variables (see Section 4.7 of Ghosal and van der Vaart (2017)). After writing this paper, we learned that this method was also recently proposed by de Carvalho et al. (2018), which is an extension of the Bayesian bootstrap method of estimation of ROC curve proposed by Gu and Ghosal (2009) for two categories.

4.1 | Notation

The notation is the same as above. Note that under the gold standard, \( L = D \), i.e., the label reflects the true disease status. Because of that, we can denote the observations from the healthy group by \( X = X_{n_0} = (1, \ldots, X_{n_0}) = (S_i : D_i = 0, i = 1, \ldots, N) \), the observations from level-1 disease by \( Y = Y_{n_1} = (Y_1, \ldots, Y_{n_1}) = (S_i : D_i = 1, i = 1, \ldots, N) \) and the observations from level-2 disease by \( Z = Z_{n_2} = (Z_1, \ldots, Z_{n_2}) = (S_i : D_i = 2, i = 1, \ldots, N) \). In the absence of verification bias, \( X, Y \) and \( Z \) are all known to us.

4.2 | Model

The model part is simplified. The only reason we are modeling the underlying distributions for three different disease groups as Dirichlet Gaussian mixture models is that we need to sample the density functions \( f_0, f_1 \) and \( f_2 \) when updating the unobserved underlying disease status \( \{ D : L = 3 \} \). Since in this case \( D = L \), we no longer need to update \( \{ D : L = 3 \} \), and thus can only focus on the distributions themselves without specifying any specific model. Thus we have

\[
X_i \overset{\text{i.i.d.}}{\sim} F_0, \quad i = 1, 2, \ldots, n_0; \quad Y_j \overset{\text{i.i.d.}}{\sim} F_1, \quad j = 1, 2, \ldots, n_1; \quad Z_k \overset{\text{i.i.d.}}{\sim} F_2, \quad k = 1, 2, \ldots, n_2.
\]

The general functional form of ROC surface is given by (3). Let \( G(t) = F_0(F_1^{-1}(t)), H(t) = F_2(F_1^{-1}(t)) \), then we have

\[
\text{ROC}_3(p_1, p_3) = \begin{cases} 
H^{-1}(1 - p_3) - G^{-1}(p_1), & \text{if } G^{-1}(p_1) \leq H^{-1}(1 - p_3) \\
0, & \text{otherwise.}
\end{cases}
\]  

(20)

As for the VUS, it is equal to the probability that three randomly selected measurements, one in each class, are in the correct order (Dreiseitl et al., 2000), i.e., \( \text{VUS} = P(X < Y < Z) \). With \( G \) and \( H \) as defined above, we have

\[
\text{VUS} = P(X < Y < Z) = P(F_0^{-1}(U_0) < F_1^{-1}(U_1) < F_2^{-1}(U_2))
\]

\[
= \int P(F_0^{-1}(U_0) < F_1^{-1}(U_1) < F_2^{-1}(U_2)|U_1 = u_1)du_1
\]

\[
= \int P(F_0^{-1}(U_0) < F_1^{-1}(U_1)|U_1 = u_1)P(F_2^{-1}(U_2)|U_1 = u_1)du_1
\]

\[
= \int P(U_0 < F_0(F_1^{-1}(u_1))|U_1 = u_1)P(U_2 > F_2(F_1^{-1}(u_1))|U_1 = u_1)du_1
\]

\[
= \int P(U_0 < G(u_1)|U_1 = u_1)P(U_2 > H(u_1)|U_1 = u_1)du_1
\]

\[
= \int G(u_1)[1 - H(u_1)]du_1,
\]  

(21)
where $U_0, U_1, U_2 \overset{i.i.d.}{\sim} U(0, 1)$; here $U(0, 1)$ stands for the uniform distribution on $(0, 1)$. Thus we only need to estimate the functions $G$ and $H$ to obtain the estimates of the ROC surface and the VUS.

### 4.3 Prior distributions and posterior computation

To conduct a Bayesian analysis, a natural choice for priors on $F_0$, $F_1$ and $F_2$ is the Dirichlet process. Let $F_0 \sim \text{DP}(M_0\xi_0)$, $F_1 \sim \text{DP}(M_1\xi_1)$, $F_2 \sim \text{DP}(M_2\xi_2)$. Then conditional on the data $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$ and $(Z_1, \ldots, Z_n)$, the posterior distribution $F_1$ is

$$F_1|\text{data} \sim \text{DP}(M_1\xi_1 + n_1\overline{F}_1).$$

Given $F_1$, the posterior distributions of $G$ and $H$ are given by

$$G|(F_1, \text{data}) \sim \text{DP}(M_0\xi_0 \circ F_1^{-1} + n_0\overline{F}_0 \circ F_1^{-1}),$$

$$H|(F_1, \text{data}) \sim \text{DP}(M_2\xi_2 \circ F_1^{-1} + n_2\overline{F}_2 \circ F_1^{-1}),$$

where $\overline{F}_0, F_1$ and $\overline{F}_2$ stand for the empirical distributions based on $X, Y$ and $Z$ respectively. However, this realization involves generating an infinite collection of random variables, which is computationally more intensive. Hence we consider the non-informative limit of the Dirichlet process by letting $M_0 \to 0$, $M_1 \to 0$ and $M_2 \to 0$. In this case, we do not even need to specify the centering measure $\xi_1, \xi_2$ and $\xi_3$. The posteriors in this case would be

$$F_1|\text{data} \sim \text{DP}(n_1\overline{F}_1), \quad G|(F_1, \text{data}) \sim \text{DP}(n_0\overline{F}_0 \circ F_1^{-1}), \quad H|(F_1, \text{data}) \sim \text{DP}(n_2\overline{F}_2 \circ F_1^{-1}),$$

known as the Bayesian bootstrap distribution.

Based on the posterior distributions above, we can obtain the algorithm for estimating the ROC surface without verification bias. Please refer to the Supplementary Materials for this algorithm. We can obtain the estimates of $\hat{G}$ and $\hat{H}$ by averaging the MCMC samples of $G$ and $H$. Plugging in the estimates to (20) and (21), we get the estimate of the ROC surface as well as the VUS.

### 5 SIMULATION

#### 5.1 With verification bias

##### 5.1.1 MAR assumption satisfied

In this simulation, we set $n_0 = n_1 = n_2 = n$, i.e., we generate $n$ data from each category and get $N = 3n$ data in total. Here we consider 6 different underlying true models:

1. Norm 1: $X \sim N(-4, 9), \quad Y \sim N(3, 16), \quad Z \sim N(5, 25), \quad \text{VUS} = 0.548.$
2. Norm 2: $X \sim N(0, 4), \quad Y \sim N(5, 9), \quad Z \sim N(8, 9), \quad \text{VUS} = 0.679.$
3. Mixnorm 1: $X \sim 0.4N(2, 25) + 0.6N(-5, 9), \quad Y \sim 0.5N(5, 16) + 0.5N(-3, 9), \quad Z \sim 0.6N(10, 9) + 0.4N(0, 25), \quad \text{VUS} = 0.462.$
4. Mixnorm 2: $X \sim N(-3, 9), \quad Y \sim 0.3N(11, 9) + 0.3N(-5, 9) + 0.4N(3, 4), \quad Z \sim 0.6N(15, 4) + 0.4N(3, 9), \quad \text{VUS} = 0.555.$
5. Logis 1: $X \sim \text{Logis}(-4, 9), \quad Y \sim \text{Logis}(3, 4), \quad Z \sim \text{Logis}(4, 36), \quad \text{VUS} = 0.452,$
   where $\text{Logis}(\mu, s^2)$ stands for the logistic distribution with location $\mu$ and scale $s$.
6. Logis 2: $X \sim \text{Logis}(0, 4), \quad Y \sim \text{Logis}(5, 9), \quad Z \sim \text{Logis}(10, 9), \quad \text{VUS} = 0.556.$

Within each model setting, we consider two different verification mechanisms which satisfy the MAR assumption proposed by Gu et al. (2014) as follows:

$$P(L_i \neq 3|S) = \begin{cases} 1, & \text{if } S > S_{(0.6\cdot n)}, \\ 0.6, & \text{otherwise}. \end{cases}$$  \hspace{1cm} (22)

$$P(L_i \neq 3|S) = \Phi(0.1 + 0.1Q_i).$$  \hspace{1cm} (23)
The first verification mechanism will give a missing rate at 36% while the second missing mechanism will give different missing rates to different model settings but will always lie in the range of 30%–45%.

The literature on ROC surface estimation is very limited for the time being. We compare our proposed Bayesian non-parametric method (denoted by DP) to Bayesian Rank Likelihood (BRL) proposed by Zhu and Ghosal (2018) and Full Imputation (FI), Mean Score Imputation (MSI), Inverse Probability Weighted (IPW) and Semi-Parametric Efficient (SPE) methods proposed by To Duc et al. (2016). We consider a multivariate logistic model to estimate the true disease status and a logistic model to estimate the verification probability when applying FI, MSI, IPW and SPE methods. Those methods are implemented using R package bcROCsurface.

The simulation results are given in Tables 1 and 2. Consider $n = 100$ and $n = 50$ and simulate 100 datasets for each setting. Both DP and BRL estimates are obtained based on 90000 Gibbs samples (100000 MCMC iterations after 10000 iterations used for burn-in). Some results are not shown either because the bias and MSE are too large or because the corresponding algorithm fails to converge.

Based on the simulation results, we find out that DP method demonstrates superiority in certain cases and is at least comparable with others in other cases. More specifically, comparing DP to BRL, DP is very competitive with BRL in Norm 1 and Norm 2 cases, where BRL is expected to have the best performance since the trinormality assumption is perfectly satisfied. DP has much better performances in Mixnorm 2 and Logis 1 compared to BRL, maybe because the model settings are too far away from the trinormality assumption. Indeed, BRL even failed to converge in logis 1 case when $n = 50$. Compared to other methods, we can see that DP outperforms FI, MSI, IPW and SPE in terms of both MSE and bias in most cases, especially in Mixnorm 1, Mixnorm 2 and Logis 1. Another observation is that FI, MSI, IPW and SPE give different results to different true verification models, while DP and BRL are more robust to the true verification models.

We can thus reach a conclusion that DP is preferred in general. While the performance of BRL depends on whether the trinormality assumption is satisfied or not, FI, MSI, IPW and SPE depend on the specification of the verification model or the true disease model. DP has a satisfactory performance in all cases considered and should be considered as a safe choice. Moreover, the ROC surfaces fitted using FI, MSI, IPW or SPE are not smooth, while DP and BRL give smooth estimates of the surface, which is a desirable property.

### 5.1.2 Departure from the MAR assumption

The only assumption we made for the DP method is the MAR assumption for missing true disease status. Thus it is important to study the performance of the DP method when this assumption is not satisfied. Here we consider the underlying distribution the same as Mixnorm 2 case, i.e., $X \sim N(-3, 9)$, $Y \sim 0.3N(11, 9)+0.3N(-5, 9)+0.4N(3, 4)$, $Z \sim 0.6N(15, 4)+0.4N(3, 9)$, VUS = 0.555. We generate $n = 100$ from each category. Since the verification no longer satisfies the MAR assumption, (6) no longer holds. Instead, we have

$$P(L_i \neq 3|S_i, D_i = d) = g_d(S_i), \quad d = 0, 1, 2.$$ (24)

The posterior distribution for disease status in this case should be adjusted:

$$P(D_i = k|S_i = t, L_i = 3) = \frac{\lambda_k g_k(t) f_k(t)}{\sum_{k=0}^{2} \lambda_k g_k(t) f_k(t) + \lambda_0 g_0(t) f_0(t) + \lambda_1 g_1(t) f_1(t) + \lambda_2 g_2(t) f_2(t)}, \quad k = 0, 1, 2;$$

see Section 4.5.

We consider two cases when the MAR assumption fails:

1. Let $g(S; \alpha, \beta) = \Phi(\alpha + \beta S)$. The verification model for healthy group is $g_0(S) = g(S; 0.15, 0.2)$; the verification model for level-1 disease is $g_1(S) = g(S; 0.1, 0.1)$; the verification model for level-2 disease is $g_2(S) = g(S; 0, 0.15)$. 

2. Let $g(S; p_1, p_2) = \begin{cases} 1, & \text{if } S > S_{(p, N)}, \\ p_2, & \text{if } S \leq S_{(p, N)}. \end{cases}$ The verification model for healthy group is $g_0(S) = g(S; 0.7, 0.3)$; the verification model for level-1 disease is $g_1(S) = g(S; 0.6, 0.4)$; the verification model for level-2 disease is $g_2(S) = g(S; 0.7, 0.4)$.

The results are shown in Table 3. The DP method has much better accuracy in terms of both bias and MSE. This shows that DP is quite robust when the MAR assumption is not satisfied.
TABLE 1 Bias (×10) and MSE (×10^2) for the estimate of the VUS, with verification bias generated using the threshold model, n = 100 and 50. (DP: Dirichlet Process, BRL: Bayesian Rank Likelihood, FI: Full Imputation, MSI: Mean Score Imputation, IPW: Inverse Probability Weighted, SPE: Semi-Parametric Efficient)

<table>
<thead>
<tr>
<th></th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Norm 1, VUS = 0.548</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Norm 2, VUS = 0.679</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>bias</td>
<td>-0.05</td>
<td>0.07</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td>-0.16</td>
<td>-0.20</td>
<td>0.09</td>
<td>0.13</td>
<td>0.08</td>
<td>0.01</td>
<td>-0.13</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.06)</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.19</td>
<td>0.16</td>
<td>0.29</td>
<td>0.27</td>
<td>0.32</td>
<td>0.37</td>
<td>0.19</td>
<td>0.16</td>
<td>0.16</td>
<td>0.25</td>
<td>0.30</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>n = 50</td>
<td>bias</td>
<td>-0.26</td>
<td>0.25</td>
<td>0.18</td>
<td>0.19</td>
<td>0.19</td>
<td>-0.06</td>
<td>-0.19</td>
<td>0.03</td>
<td>0.11</td>
<td>0.07</td>
<td>0.07</td>
<td>-0.07</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.08)</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.30</td>
<td>0.34</td>
<td>0.57</td>
<td>0.53</td>
<td>0.65</td>
<td>0.67</td>
<td>0.29</td>
<td>0.29</td>
<td>0.45</td>
<td>0.43</td>
<td>0.52</td>
<td>0.60</td>
<td>(0.04)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.12)</td>
</tr>
</tbody>
</table>

5.2 Without verification bias

For simplicity, the simulation results for estimating ROC surface without verification bias are shown in the Supplementary Materials. The Bayesian bootstrap (BB) method has overall satisfactory results compared to four other methods. Note that the speed of the BB method is a lot faster compared with the DP method. In fact, if we run a dataset with 50 observations in each category using the method under verification bias (implemented in R), it takes approximately 30 minutes. The simplified method without verification bias (implemented in MatLab), however, finish the computation in about 1 second.

6 REAL DATA ANALYSIS

6.1 Epithelial ovarian cancer

Epithelial ovarian cancer (EOC) is one of the most lethal cancers in adult women. CA125 and Human Epididymis protein 4 (HE4) are the most powerful and widely used biomarkers in diagnosing EOC. They are actually the only two markers approved by the FDA for detecting the disease to date. There are a lot of studies which compare the performance of these two biomarkers in detecting EOC. Zheng and Gao (2012) and Hamed et al. (2013) claim that HE4 is a more powerful diagnostic tool while Cramer et al. (2011) suggest that CA125 is the best biomarker. However, most studies did not consider the performance of
The VUS for CA125 is estimated to be approximately 0.05. In this case, we can apply both DP and BRL methods. Otherwise, BRL method may not be very appropriate.

The dataset we considered is publicly available. This is the experiment data taken from the SPORE/Early Detection Network/Prostate, Lung, Colon, and Ovarian Cancer Ovarian Validation Study. The data contains 703 control cases (156 controls with benign disease, 471 general population controls and 76 controls unsure with benign disease or not), 72 early-stage cases, 78 late-stage cases, and 87 cases unverified. We will use this data to compare CA125 and HE4 by applying our method along with other methods which can treat verification bias.

Table 4 summarizes the estimates of the VUS for CA125 and HE4. The DP and BRL methods are both based on 80000 iterations (100000 iterations with the first 20000 iterations as burn-in). Before applying BRL methods, we did a quick visual check on the data to see if it satisfies the trinormality assumption. The details of the method for checking the assumption is described in the Supplementary Materials along with the result. It turns out that this data actually satisfies the trinormality assumption quite well. In this case, we can apply both DP and BRL method. Otherwise BRL method may not be very appropriate. The VUS for CA125 is estimated to be approximately 0.36, while the estimate given by FI is slightly larger than others. The VUS

<table>
<thead>
<tr>
<th></th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VUS = 0.548</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>-0.11</td>
<td>0.10</td>
<td>0.17</td>
<td>0.18</td>
<td>0.09</td>
<td>&lt; 0.01</td>
<td>-0.13</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>&lt; 0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.18</td>
<td>0.18</td>
<td>0.30</td>
<td>0.29</td>
<td>0.30</td>
<td>0.30</td>
<td>0.13</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>n = 50</td>
<td>-0.23</td>
<td>0.26</td>
<td>0.14</td>
<td>0.15</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.36</td>
<td>0.21</td>
<td>0.28</td>
<td>0.22</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.46</td>
<td>0.42</td>
<td>0.51</td>
<td>0.47</td>
<td>0.53</td>
<td>0.55</td>
<td>0.36</td>
<td>0.41</td>
<td>0.35</td>
<td>0.33</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.05)</td>
</tr>
</tbody>
</table>

Logis 1, VUS = 0.452 Logis 2, VUS = 0.556

<table>
<thead>
<tr>
<th></th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VUS = 0.548</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 100</td>
<td>-0.13</td>
<td>0.38</td>
<td>0.39</td>
<td>0.54</td>
<td>-0.01</td>
<td>-0.06</td>
<td>-0.10</td>
<td>0.15</td>
<td>0.14</td>
<td>0.16</td>
<td>-0.05</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.19</td>
<td>0.28</td>
<td>0.44</td>
<td>0.55</td>
<td>0.44</td>
<td>0.57</td>
<td>0.22</td>
<td>0.23</td>
<td>0.16</td>
<td>0.20</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.09)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>n = 50</td>
<td>-0.16</td>
<td>-</td>
<td>0.45</td>
<td>0.60</td>
<td>0.07</td>
<td>-0.05</td>
<td>-0.41</td>
<td>0.17</td>
<td>0.18</td>
<td>0.21</td>
<td>-0.03</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.13)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.38</td>
<td>-</td>
<td>0.85</td>
<td>0.95</td>
<td>0.87</td>
<td>1.73</td>
<td>0.47</td>
<td>0.54</td>
<td>0.36</td>
<td>0.38</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.12)</td>
<td>(0.13)</td>
<td>(0.11)</td>
<td>(0.90)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

TABLE 3 Bias ($\times 10$) and MSE ($\times 10^2$) for the estimate of the VUS, with verification bias departing from NMAR model, $n = 100$. (DP: Dirichlet Process, BRL: Bayesian Rank Likelihood, FI: Full Imputation, MSI: Mean Score Imputation, IPW: Inverse Probability Weighted, SPE: Semi-Parametric Efficient)

<table>
<thead>
<tr>
<th></th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
<th>DP</th>
<th>BRL</th>
<th>FI</th>
<th>MSI</th>
<th>IPW</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMAR model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>−0.01</td>
<td>0.53</td>
<td>0.60</td>
<td>0.51</td>
<td>−0.60</td>
<td>−0.69</td>
<td>0.29</td>
<td>0.36</td>
<td>0.80</td>
<td>0.75</td>
<td>−0.20</td>
<td>−0.65</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.07)</td>
<td>(0.10)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.22</td>
<td>0.44</td>
<td>0.54</td>
<td>0.47</td>
<td>0.88</td>
<td>1.42</td>
<td>0.20</td>
<td>0.31</td>
<td>0.80</td>
<td>0.75</td>
<td>0.45</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.11)</td>
<td>(0.24)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.08)</td>
<td>(0.07)</td>
<td>(0.18)</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 4 The VUS estimates for CA125 and HE4.

<table>
<thead>
<tr>
<th></th>
<th>CA125</th>
<th>HE4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimated VUS</td>
<td>sd</td>
</tr>
<tr>
<td>DP</td>
<td>0.365</td>
<td>0.056</td>
</tr>
<tr>
<td>BRL</td>
<td>0.369</td>
<td>0.029</td>
</tr>
<tr>
<td>FI</td>
<td>0.427</td>
<td>0.030</td>
</tr>
<tr>
<td>MSI</td>
<td>0.365</td>
<td>0.034</td>
</tr>
<tr>
<td>IPW</td>
<td>0.356</td>
<td>0.035</td>
</tr>
<tr>
<td>SPE</td>
<td>0.354</td>
<td>0.035</td>
</tr>
</tbody>
</table>

for HE4 is estimated to be around 0.44, with a slightly smaller estimate given by BRL. The estimates given by different methods are quite similar overall. The standard deviation for DP is a bit larger than other estimators, so its corresponding confidence interval (C.I.) is also wider than others.

It seems that the VUS is slightly larger for HE4 but it is hard to conclude which biomarker is better, since the confidence intervals intersect with one another. To better compare the fitted ROC surface using all those methods, we also plot out the ROC surface.

We can see from Figures 1 and 2 that the DP and BRL methods generate smooth surfaces while others do not. Judging from the plots, the ROC surfaces estimated by MSI, IPW and SPE are very much alike and the ROC surfaces generated by DP are like a smoother version of those estimated surfaces. The surfaces estimated by BRL, on the other hand, seem to have smoothed out too many details, and for some reasons, look more like the surfaces generated by FI. The figures give us a rough impression that HE4 has better diagnosis ability than CA125.

6.2 Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is one of the most common type of liver cancer. It is the fifth most common tumor worldwide and the second most common cause of cancer-related death. HCC usually occurs in people with chronic liver diseases, for example cirrhosis caused by hepatitis B or hepatitis C.

Albumin is an important serum protein in human body. Studies have shown that serum albumin levels have prognostic significance in HCC, with lower serum albumin levels having significantly larger maximum tumor diameters. In fact, serum albumin level is already being used as a criterion for the HCC staging systems.

Here, we want to apply our method to assess the ability of serum albumin level in distinguishing different stages of HCC. The dataset we are using is publicly available on Kaggle. The dataset contains 159 records in total with 2 patients in the very early stage, 29 patients in the early stage, 39 patients in the intermediate stage, 53 patients in the advanced stage, 37 patients in the terminal stage and 5 patients unverified. We combine the very early stage and early stage, intermediate stage and advanced stage to get three outcomes in total. The dataset contains the serum albumin levels for every patient as well as a covariate, gender. There

3https://www.kaggle.com/mrsantos/hcc-dataset/version/5
We want to utilize the covariate information when we are assessing this biomarker. We then use the extension of our proposed method to incorporate covariates when estimating the ROC surface. The other methods for estimating the ROC surface under verification bias cannot incorporate covariates, so we only use the DP method for illustration. The results is shown in Table 5. We calculate the VUS for the covariate-specific ROC surface as well as for the overall ROC surface.

**FIGURE 1** Estimated ROC surfaces for CA125

**TABLE 5** The VUS estimates for serum albumin

<table>
<thead>
<tr>
<th></th>
<th>estimated VUS</th>
<th>sd</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.390</td>
<td>0.117</td>
<td>[0.161, 0.619]</td>
</tr>
<tr>
<td>Male</td>
<td>0.407</td>
<td>0.072</td>
<td>[0.266, 0.548]</td>
</tr>
<tr>
<td>Overall</td>
<td>0.406</td>
<td>0.067</td>
<td>[0.275, 0.536]</td>
</tr>
</tbody>
</table>

are 29 females and 130 males in this dataset. (The male-to-female incidence ratio varies between 2:1 to 4:1 in the population.[33])
FIGURE 2 Estimated ROC surfaces for HE4

The ROC surface plots are shown in Figure 2. We can see that the ROC surfaces for female and male are substantially different. The overall ROC surface looks more like the ROC surface for male which make sense since the ratio of male patients is much larger than the ratio of female patients.

Our method confirms that serum albumin has a prognostic ability to identify the stage of HCC. Its VUS for distinguishing different stages of HCC is around 0.406. However, serum albumin alone does not appear to be sufficient for staging the HCC.

7 | DISCUSSION

Our methods can be easily extended to estimate the ROC hypersurface which is defined when there are more than three categories to be classified. This is originally defined by Scrufteld (1996)\textsuperscript{230}

\[
\{(F_1(c_1), F_2(c_2) - F_2(c_1), F_3(c_3) - F_3(c_2), \ldots, F_{d-1}(c_{d-1}) - F_{d-1}(c_{d-2}), 1 - F_d(c_{d-1})) : c_1 < c_2 < \ldots < c_{d-1}\}
\]

for a $d$-dimensional ROC hypersurface, where $F_i$ denote the true distribution for measurements from the $i$th category, $i = 1, 2, \ldots, d$. Similar to the AUC and the VUS, he proposed using hypervolume under the ROC manifold (HUM) as a measure
of the performance of a classifier, and HUM can be calculated as $\text{HUM} = P(X^{(1)} < X^{(2)} < \cdots < X^{(d)})$, where $X^{(i)}$ denote the measurement from the $i$th category. With the knowledge of the true distributions $F_1, F_2, \ldots, F_d$, it can also be calculated as $\text{HUM} = P(F^{-1}_1(U_1) < F^{-1}_2(U_2) < \cdots < F^{-1}_d(U_d))$, where $U_1, \ldots, U_d \sim \text{U}(0, 1)$. Through some simple modifications, our methods introduced above can be extended to estimate $F_1, F_2, \ldots, F_d$ with the measurements from $d$ categories. We can then give an estimate of the ROC hypersurface as well as the HUM.

References


7. Carvalho dVI, Carvalho dM, Branscum A. Bayesian Bootstrap Inference for the ROC Surface. Stat 2018; 7. doi: https://doi.org/10.1002/sta4.211