Making Diagnoses Based on Multiple Diagnostic Tests Without a Gold Standard

Jingyang Zhang, Ying Zhang, Kathryn Chaloner and Jack T Stapleton

Abstract

When there is no "gold standard" available, it is common to reconcile information from multiple imperfect diagnostic tests in order to obtain more accuracy. In this paper, we generalize the linear discriminant method and the optimal risk score method to accommodate the situation that is lack of a "gold standard". We also study an alternative sequential diagnostic method which does not require all tests to be applied to each subject. All the methods are developed under some parametric distributional assumptions. A mixture of two multivariate normal distributions is used to fit the unclassified data and the optimal diagnostic rule for each method is derived based on the fitted model. We provide the numerical implementation of all methods. Asymptotic results of statistical inferences about the methods are also given. Simulation studies are carried out to compared the methods and the illustration with a real-life data set is included.

1 Introduction

In medical applications, it is common to use multiple diagnostic tests or diagnostic markers for the detection and diagnosis of an illness. Diagnostic accuracy is generally assessed by the receiver operating characteristic (ROC) curve, which is defined as the plot of "sensitivity" against 1-"specificity" over all possible cutoff values (Zhou et al., 2002; Pepe, 2003). Different tests may be compared by the area under ROC curve (AUC) (Hanley and McNeil, 1982, 1983; McClish, 1987; DeLong et al., 1988), or the sensitivity at a fixed common specificity (Greenhouse and Mantel, 1950; Linnet, 1987).

When multiple tests are involved, the performance of the diagnosis may be improved by combining several tests as a new composite diagnostic test since different tests may be sensitive to different aspects of case. For example, detection of a particular antibody in a biosample may be best determined by application of tests with more than one specificity and/or sensitivity. There have been several methods proposed to combine multiple tests. A linear discriminant method is proposed by Su and Liu (1993) for the scenario that the values of diagnostic markers form multivariate normal distributions for both case and non-case (control) populations. The linear discriminant analysis identifies the

optimal linear combination of the tests that maximises the sensitivity over the entire specificity range uniformly. Alternatively, McIntosh and Pepe (2002) use a risk score derived from the Neyman-Pearson lemma (Neyman and Pearson, 1933) to combine tests. The risk score can be estimated by binary regression models without assuming the distributions of the diagnostic markers and it has been justified optimal to discriminate case from non-case subjects (McIntosh and Pepe, 2002).

The aforementioned methods are developed for the situation in which the true event status is known for study subjects, or a perfect "gold standard" as it is often called (Zhou et al., 2002; Pepe, 2003) exists. However, the "gold standard" is not always available, and this leads to situations for which it is difficult or impossible to establish a definite diagnosis. In a motivating example of antibody tests for a common, nonpathogenic flavivirus (GB virus C or GBV-C), we have two enzyme linked immunosorbent assay (ELISA) tests (Tacke et al., 1997; McLinden et al., 2006) that were applied to 100 unique blood samples in order to detect the presence of antibody against the major envelope glycoprotein of GBV-C (E2) in blood samples. In immune competent individuals, GBV-C infection is frequently cleared by the immune system, and following clearance, antibody to E2 is detected (Tacke et al., 1997). Among people with HIV infection, persistent GBV-C co-infection is associated with prolonged survival in HIV patients (Lefrère et al., 1999; Xiang et al., 2001; Tillmann et al., 2001; Williams et al., 2004; Van der Bij et al., 2005; Zhang et al., 2006). Subjects without GBV-C who have GBV-C E2 antibodies, however, also appear to survive longer from HIV infection than those who do not have either active infection or evidence of prior infection (E2 antibodies) (Tillmann et al., 2001; Williams et al., 2004). Currently, for this antibody, there is no commercial or validated test available. The data set only contains the results from the two "in-house" ELISA tests developed to detect antibodies to GBV-C E2. One hundred independent blood samples were analysed by both methods. One of the study aims is to develop a diagnostic test using the results from two ELISA tests. The linear method and the risk score method described above cannot be applied directly since no definite classification for the E2 antibodies is available in the data.

In this article, we generalize Su-Liu's linear discriminant method and MnIntosh-Pape's risk score method to the situation that no "gold-standard" exists. The methods are developed by fitting a two-term multivariate normal mixture model to unclassified data on the results of diagnostic tests with two terms corresponding to case and non-case respectively (Hui and Zhou, 1998). The parameters of multivariate normal distributions and the event prevalence are estimated using maximum likelihood estimation with EM algorithm. Then Su-Liu's test and MnIntosh-Pepe's test are easily evaluated as the tests can be expressed as some functions of the parameters of the multivariate normal mixture distribution. In addition, we also develop a diagnostic method in a sequential fashion in which the second test is implemented only if the first test does not result in a positive diagnosis. The idea of combining tests in a sequence was first discussed by Kraemer (1992), and Thompson (2003) discussed theoretical accuracy of a sequence of tests. This method is practically sound when people know one test

is definitely more sensitive to the event than the other and do not wish to conduct multiple diagnostic tests either due to complication of the diagnostic tests or high cost associated with the tests.

The rest of the article is organised as follows. Section 2 presents the three methods for combining multiple tests without "gold standard". Section 3 describes the numerical implementation of the methods. Section 4 discusses statistical inferences about the methods. Section 5 provides simulation studies to compare the methods and illustrates them using the data from the motivating ELISA example. Some concluding remarks are included in Section 6 and the technical proofs for statistical inference are outlined in the appendix.

2 Methods

For the simplicity in illustration, suppose that there are two quantitative diagnostic tests on each subject and for each test, a greater value of the test result indicates a higher chance of case. Denote X_i as the random variable representing the result from test i for $i = 1, 2$ and D as the random variable indicating the case presence, with $D = 1$ meaning case present and $D = 0$ meaning case absent. Moreover, F_1 and F_0 are the joint distribution functions of $\mathbf{X} = (X_1, X_2)$ in the $D = 1$ and $D = 0$ population, respectively, and f_1 and f_0 are the corresponding probability density functions.

2.1 The Optimal Linear Composite Method

Suppose $\mathbf{X}$ is normally distributed under both $D = 1$ and $D = 0$ with different model parameters, i.e., $\mathbf{X}|D = 1 \sim N(\mu_1, V_1)$ and $\mathbf{X}|D = 0 \sim N(\mu_0, V_0)$. Su and Liu (1993) considered a linear combination $U = \mathbf{a}^T\mathbf{X}$ of the two diagnostic tests as a composite diagnostic test. Under the normality assumption, the coefficients $\mathbf{a}$ corresponding to the optimal linear composite test, which provides the highest sensitivity uniformly at any specificity among all possible linear composite tests, are given by (1).

$$\begin{aligned}\mathbf{a}_0 &\propto V_0^{-1/2}\left(I + V_0^{-1/2}V_1V_0^{-1/2}\right)^{-1}V_0^{-1/2}(\mu_1 - \mu_0)\\ &= (V_0 + V_1)^{-1}(\mu_1 - \mu_0),\end{aligned} \tag{1}$$

where I is the 2×2 identity matrix.

The ROC curve and the AUC of the optimal linear-composite test are easily calculated as

$$ROC(u) = \Phi\left(\frac{\mathbf{a}_0^T(\mu_1 - \mu_0) + \Phi^{-1}(u)\sqrt{\mathbf{a}_0^TV_0\mathbf{a}_0}}{\sqrt{\mathbf{a}_0^TV_1\mathbf{a_0}}}\right), \tag{2}$$

$$AUC = \Phi\left(\sqrt{(\mu_1 - \mu_0)^T(V_0 + V_1)^{-1}(\mu_1 - \mu_0)}\right), \tag{3}$$

where Φ is the cumulative distribution function of the standard normal distribution.

For diagnoses at a given specificity p_0, the threshold for the optimal linear composite test is $\mathbf{a}_0^T\mu_0 + \Phi^{-1}(p_0)\sqrt{\mathbf{a}_0^T V_0 \mathbf{a}_0}$, and the corresponding sensitivity is

$$\text{sen}_A = \Phi\left(\frac{\mathbf{a}_0^T(\mu_1 - \mu_0) - \Phi^{-1}(p_0)\sqrt{\mathbf{a}_0^T V_0 \mathbf{a}_0}}{\sqrt{\mathbf{a}_0^T V_1 \mathbf{a}_0}}\right) \tag{4}$$

according to (2). Su-Liu's method can be automatically applied to the situation when no "gold standard" exists, as long as the model parameters can be consistently estimated.

2.2 The Optimal Risk Score Composite Method

McIntosh and Pepe (2002) developed a composite diagnostic test in the framework of hypothesis testing in which $D = 0$ and $D = 1$ represent the null and alternative hypotheses, respectively. The decision rule to classify D as one is analogous to the rule for rejecting the null hypothesis in favor of the alternative. With this analogy, type I error corresponds to the false positive rate (FPR) and the power corresponds to the true positive rate (TPR). Based on the Neyman-Pearson lemma (Neyman and Pearson, 1933) for the likelihood ratio test, it is easily established that the decision rule defined as

$$\begin{cases} \text{D=1} & \text{if } LR(\mathbf{X}) > c(p_0); \\ \text{D=0} & \text{otherwise,} \end{cases}$$

where $LR(\mathbf{X}) = \Pr(\mathbf{X}|D=1)/\Pr(\mathbf{X}|D=0)$ and $c(p_0)$ satisfies that $\Pr(LR(\mathbf{X}) > c(p_0)|D=0) = 1 - p_0$. The rule is the uniformly most sensitive (UMS) diagnostic test among all tests with FPR$= 1 - p_0$.

To evaluate the risk score $LR(\mathbf{X})$, one needs to estimate the distributions of the diagnostic markers X in advance. Using Bayes rule, McIntosh and Pepe (2002) demonstrated that the decision rule can be equivalently postulated based on an alternative risk score $p(\mathbf{X}) = \Pr(D=1|\mathbf{X})$ as

$$\begin{cases} \text{D=1} & p(\mathbf{X}) = \frac{LR(\mathbf{X})q}{LR(\mathbf{X})q+1} > c^*(p_0); \\ \text{D=0} & \text{otherwise,} \end{cases}$$

where $q = \Pr(D=1)/\Pr(D=0)$, and $\Pr(p(\mathbf{X} > c^*(p_0)|D=0) = 1 - p_0$. This risk score can be easily estimated via logistic regression $logit(p(\mathbf{X})) = \beta_0 + h(\beta, \mathbf{X})$ without the knowledge of the markers' distribution. But MnIntosh-Pepe's risk score is only applicable when a "gold standard" exists so that subjects can be classified without error.

When the diagnostic markers' distributions are estimable, the aforementioned risk score based composite test is also valid even though no "gold standard" is available. For continuous diagnostic markers, we use the original risk score

$$LR(\mathbf{X}) = f_1(\mathbf{X})/f_0(\mathbf{X}).$$

The threshold $c(p_0)$ under the specificity p_0 is in fact the p_0 percentile of $LR(\mathbf{X})$ evaluated for the $D = 0$ group. Let H_1 and H_0 denote the distribution function of $LR(X)$ under $X \sim F_1$ and $X \sim F_0$, respectively. Then the sensitivity for the threshold $c(p_0)$ is simply

$$\text{sen}_B = 1 - H_1\left(H_0^{-1}(p_0)\right). \tag{5}$$

To determine the decision rule and its sensitivity corresponding to a fixed specificity, one just need to have some consistent estimates for the distribution function H_1 and the quintile of H_0.

2.3 The Optimal Sequential Composite Method

The composite tests described above require subjects to undergo all the diagnostic tests that may not be desirable in practice in view of potential risks associated with the diagnostic tests or the excessive financial burden with extra tests, particularly if the tests are all expensive. In many medical diagnostic procedures, one may start with the most sensitive test among all available diagnostic tools and continue to the second test only if the diagnosis based on the first test is not conclusive. This practical diagnostic procedure motivates us to design an optimal composite test in a sequential fashion, to study its statistical properties, and to compare its performance with the other two composite tests.

Suppose Test 1 is superior to Test 2 judged by a greater value of AUC. The decision rule driven by the sequential composite test is determined by a pair of cut-off values (C_1, C_2) such that:

1. if $X_1 > C_1$, then this subject is classified as positive for the case; else,
2. if $X_2 > C_2$, then classified as positive;
3. otherwise, classified as negative.

Figure 1 depicts the classification partition in diagnostic test domain with the 2 cut-off sequential composite test.

Given the cut-off (C_1, C_2), the sensitivity and specificity for evaluating this composite test can be expressed as follows:

$$\begin{aligned}
\text{Sensitivity} &= \Pr\left(\text{Positive classification}|\text{event present}\right) \\
&= \Pr\left(X_1 > C_1 | D = 1\right) + \Pr\left(X_1 \le C_1,\ X_2 > C_2 | D = 1\right) \\
&= 1 - F_1(C_1, C_2). \qquad (6) \\
\text{Specificity} &= \Pr\left(\text{Negative classification}|\text{event absent}\right) \\
&= \Pr\left(X_1 \le C_1,\ X_2 \le C_2 | D = 0\right) \\
&= F_0(C_1, C_2). \qquad (7)
\end{aligned}$$

We are searching for the optimal sequential composite test in the sense that it achieves the maximum sensitivity among all the sequential composite tests

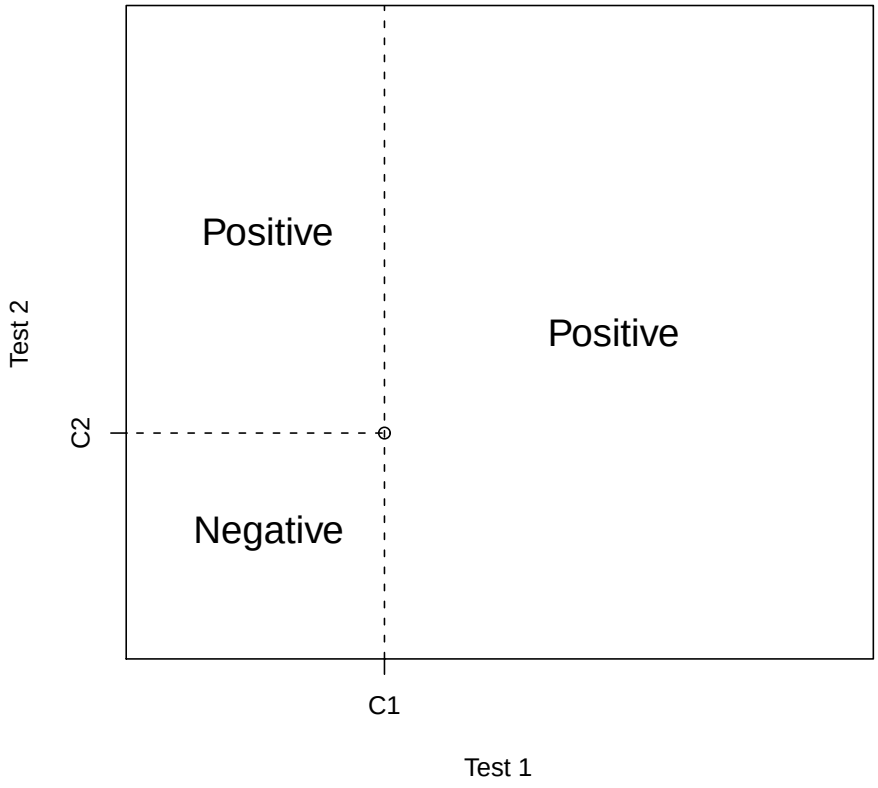

Figure 1: Illustration of the sequential composite test.

whose specificity is fixed at p_0 . Based on (6) and (7), this task is then converted to a constrained non-linear optimasation problem:

$$\min_{F_0(C_1,C_2)=p_0} F_1(C_1, C_2). \tag{8}$$

An efficient algorithm for finding the optimal (C_1, C_2) in (8) is essential in the development of this sequential method.

3 Computation of the Methods

3.1 MLE of Multivariate Normal Mixture Model

Suppose we have a sample of diagnostic markers $X_1, X_2, \cdots, X_n$ that are assumed to be independently and identically distributed copies of $\mathbf{X}$ with distribution function F. The implementation of all the foregoing methods requires estimation of F_1 and F_0 from observed data in the first place. Here we follow the set-up of Su and Liu (1993) for the distribution of the diagnostic markers $\mathbf{X}$, i.e. $\mathbf{X}|D = 1 \sim F_1 \equiv N(\mu_1, V_1)$ and $\mathbf{X}|D = 0 \sim F_0 \equiv N(\mu_0, V_0)$. We adopted the mixture distribution of F_1 and F_0 to model the observed data, that is

$$F_\theta(\cdot) = \pi F_{1,\theta_\mathbf{1}}(\cdot) + (1 - \pi)F_{0,\theta_\mathbf{0}}(\cdot), \tag{9}$$

where π is an unknown parameter indicating the mixture proportion, or equivalently, the case prevalence, and $\theta = (\pi, \theta_1, \theta_0) = (\pi, (\mu_1, V_1), (\mu_0, V_0))$ denotes the model parameters. The log-likelihood of the observed data can be expressed as:

$$\begin{aligned} l(\theta) &= \sum_{k=1}^{n} \log f_\theta(X_{1k}, X_{2k}) \\ &= \sum_{k=1}^{n} \log \left[\pi f_{1,\theta_\mathbf{1}}(X_{1k}, X_{2k}) + (1 - \pi)f_{0,\theta_\mathbf{0}}(X_{1k}, X_{2k})\right]. \end{aligned}$$

In principal, the ML estimates of the parameters $\hat{\theta}_n$ can be estimated by directly maximising the log likelihood $l(\theta)$, it is found that this approach is not numerically stable. We note that if the "gold standard" does exist so that the exact memberships $\mathcal{D} = (D_1, \ldots, D_n)$ are known, the log likelihood for the augmented data $\{(X_1, D_1), \cdots, (X_n, D_n)\}$ is given by

$$l_a(\theta) = \sum_{k=1}^{n} D_k \log \pi f_{1,\theta_1}(X_{1k}, X_{2k}) + (1 - D_k) \log(1 - \pi)f_{0,\theta_0}(X_{1k}, X_{2k})$$

and

$$Pr(D_k = 1|(X_1, \cdots, X_n); \theta) = \frac{\pi f_{1,\theta_1}(X_{1k}, X_{2k})}{\pi f_{1,\theta_1}(X_{1k}, X_{2k}) + (1 - D_k) \log(1 - \pi)f_{0,\theta_0}(X_{1k}, X_{2k})}.$$

Hence the ML estimates of the model parameters $\hat{\theta}_n$ are easily computed using the EM algorithm (Dempster et al., 1977) due to its numerical stability and algorithmic convenience for this problem.

3.2 Computation of the Optimal Linear Composite Test

Obtaining $\hat{\theta}_n$, the estimates of the coefficients of the optimal linear composite test are directly computed by plugging $\hat{\theta}_n$ in (1) as

$$\hat{\mathbf{a}}_0 \propto (\hat{V}_0 + \hat{V}_1)^{-1}(\hat{\mu}_1 - \hat{\mu}_0), . \tag{10}$$

Accordingly, for the fixed specificity p_0, the threshold is estimated by $\hat{\mathbf{a}}_0^T\hat{\mu}_0 + \Phi^{-1}(p_0)\sqrt{\hat{\mathbf{a}}_0^T\hat{V}_0\hat{\mathbf{a}}_0}$, and the sensitivity by

$$\widehat{\text{sen}}_A = \Phi\left(\frac{\hat{\mathbf{a}}_0^T(\hat{\mu}_1 - \hat{\mu}_0) - \Phi^{-1}(p_0)\sqrt{\hat{\mathbf{a}}_0^T\hat{V}_0\hat{\mathbf{a}}_0}}{\sqrt{\hat{\mathbf{a}}_0^T\hat{V}_1\hat{\mathbf{a}}_0}}\right). \tag{11}$$

3.3 Computation of the Optimal Risk Score Composite Test

With the multivariate normal distribution assumption for the diagnostic markers $\mathbf{X}$, the distribution functions of the risk score $LR(\mathbf{X})$ under $D = 0$ and $D = 1$ can be estimated empirically by drawing a random sample from $F_{0,\hat{\theta}_{0n}}$ and $F_{1,\hat{\theta}_{1n}}$, respectively. The estimation of the threshold for the decision rule and its corresponding sensitivity can be computed in the following subsequent steps:

- First, draw a random sample of $V_{n,1}, \ldots, V_{n,m}$ from $F_{0,\hat{\theta}_{0n}}$, then form a sample of $\{Z_{n,i} = LR_{\hat{\theta}_n}(V_{n,i}) = f_{1,\hat{\theta}_{1n}}(V_{n,i})/f_{0,\hat{\theta}_{0n}}(V_{n,i}),\ i = 1, \ldots, m\}$
- Compute the empirical distribution of

  $$\mathbb{H}_{0,\hat{\theta}_n,m}(Z) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_{[Z_{n,i} \le Z]}$$

  to estimate the distribution function of H_0. Then the threshold $\hat{c}(p_0)$ can be estimated by the sample p_0 percentile of $Z_{n,1}, \ldots, Z_{n,m}$, denoted as $\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(p_0)$;
- To estimate the sensitivity for the threshold $\hat{c}(p_0)$, draw a random sample of $W_{n,1}, \ldots, W_{n,m}$ from $F_{1,\hat{\theta}_{1n}}$, then form a sample of $\{Y_{n,i} = LR_{\hat{\theta}_n}(W_{n,i}),\ i = 1, \ldots, m\}$
- Similarly, compute the empirical distribution of

  $$\mathbb{H}_{1,\hat{\theta}_n,m}(Y) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_{[Y_{n,i} \le Y]},$$

  and estimate the sensitivity by $\widehat{\text{sen}}_B = 1 - \mathbb{H}_{1,\hat{\theta}_n,m}(\hat{c}(p_0))$, denoted as, $1 - \mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(p_0)\right)$.

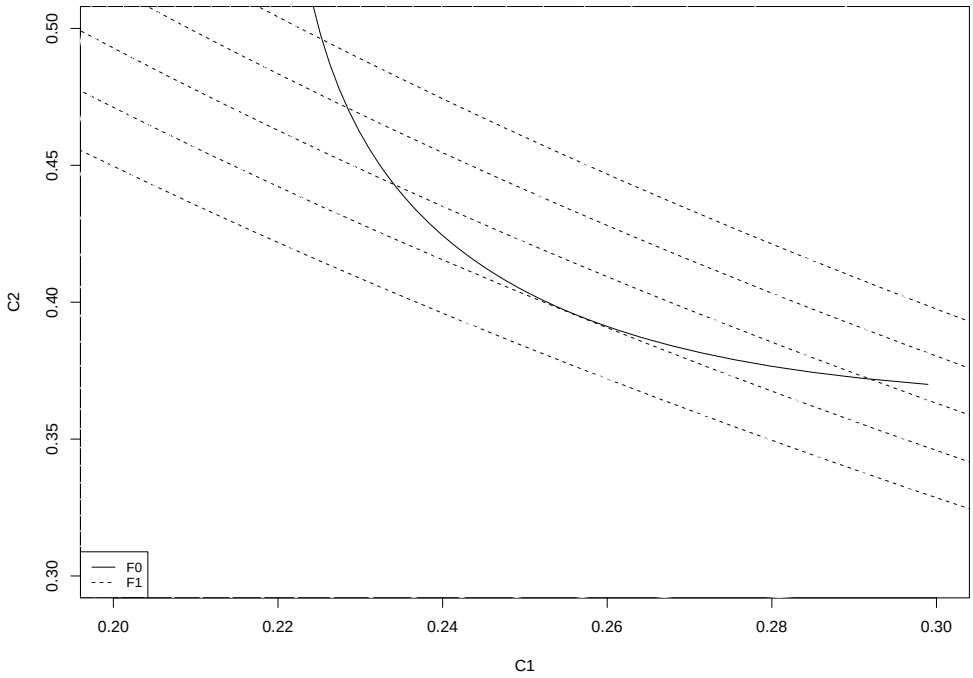

Figure 2: Illustration of the search for the optimal (C_1, C_2). The solid line is the contour curve of F_0 for the non-case population at a given specificity p_0 , and the dashed lines are the contour lines of F_1 for the case population.

3.4 Computation of the Optimal Sequential Composite Test

Under the normality assumption, the feasible set of (C_1, C_2) defined by a given specificity $F_0(C_1, C_2) = p_0$ constitutes a convex contour curve (Tihansky, 1972). When the diagnostic markers are more variant for the case subjects, it is expected that the contour given by $F_1(C_1, C_2) = t$ is also convex but with less curvature and moves towards the origin of (C_1, C_2) domain as t decreases. The optimisation problem (8) can be illustrated geometrically in Figure 2. As seen in the figure, the constrained optimal value t corresponds to the contour that touches the contour of $F_0(C_1, C_2) = p_0$. The optimal threshold (C_1, C_2) for the decision rule is simply the tangent point of the two contour lines and can be uniquely determined. Therefore, the original optimisation problem (8) is converted to solving the system of bivariate nonlinear equations (12) for the tangent point of the contour lines of F_1 and F_0.

The first equation represents the constraint given by the fixed specificity and

the second equation reflects that the two contour lines have the same gradient at the tangent point. The Newton-Raphson method with the step-halving line search procedure is utilised to solve the system (12).

$$G(\mathbf{C},\theta)=\begin{cases} F_{0,\theta_0}(C_1,C_2)=p_0 \\ \dfrac{\partial F_{1,\theta_1}}{\partial C_1}(C_1,C_2)\dfrac{\partial F_{0,\theta_0}}{\partial C_2}(C_1,C_2)-\dfrac{\partial F_{1,\theta_1}}{\partial C_2}(C_1,C_2)\dfrac{\partial F_{0,\theta_0}}{\partial C_1}(C_1,C_2)=0. \end{cases} \tag{12}$$

Let $\hat{\mathbf{C}}_n=(\hat{C}_{1n},\hat{C}_{2n})$ denote the solution of (12) with the ML estimates $\hat{\theta}_n=(\hat{\theta}_{1n},\hat{\theta}_{0n})$, then the sensitivity is estimated by $\widehat{\text{sen}}_C=1-F_{1,\hat{\theta}_{1n}}(\hat{C}_{1n},\hat{C}_{2n})$.

4 Statistical Inference

Suppose θ are the true model parameters under the mixture of bivariate normal distribution. By the MLE properties (van der Vaart, 2000), it is known that as $n\to\infty$, $\hat{\theta}_n\to_P\theta$, and

$$\sqrt{n}\left(\hat{\theta}_n-\theta\right)\to_d N(0,\mathcal{I}^{-1}),$$

where $\mathcal{I}$ is the Fisher information matrix given by $-E\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\middle|\theta\right]$.

For the optimal linear composite method, the estimated sensitivity $\hat{\text{sen}}_A$ as given in (11), is a continuous function of $\hat{\theta}_n$, hence it is consistent and asymptotically normally distributed by the ordinary continuous mapping theorem and the delta method.

For the optimal risk score method, the true sensitivity under the mixture bivariate normal distribution with the true model parameters θ is given by

$$\text{sen}_B=1-H_{1,\theta}\left(H_{0,\theta}^{-1}(p_0)\right).$$

The estimated sensitivity $\widehat{\text{sen}}_B$ can be shown consistent and asymptotically normal using the result of Theorem 1. The proof of this theorem is outlined in Appendix A.

Theorem 1. *If m satisfies that $m/n\to\infty$ as $n\to\infty$, and furthermore, if $H_{1,\theta}(x)$ and $H_{0,\theta}(x)$ are continuously differentiable with respect to both θ and x and their first derivatives are bounded, then for any fixed $q_0\in(0,1)$, $\sqrt{n}\left[\widehat{\text{sen}}_B-\text{sen}_B\right]$ converges weakly to a normal distribution with mean 0 and variance given by* (18).

Let $\mathbf{C}_0=(C_{10},C_{20})$ denote the solution of the system (12) under the true parameters θ, then the true sensitivity is $\text{sen}_C=1-F_{1,\theta}(C_{10},C_{20})$. The estimated sensitivity $\hat{\text{sen}}_C$ is also consistent and asymptotically normal under the mild condition (11) given in Theorem 2. The proof of the theorem is also outlined in Appendix B

Theorem 2. *If F_0 and F_1 are continuously differentiable with respect to $\mathbf{C} = (C_1, C_2)$ and θ and satisfy the following inequality (13) at C_0 and θ,*

$$\begin{aligned}&\left[\frac{\partial^2 F_1}{\partial C_1 \partial C_2}\frac{\partial F_0}{\partial C_2} + \frac{\partial F_1}{\partial C_1}\frac{\partial^2 F_0}{\partial {C_2}^2} - \frac{\partial^2 F_1}{\partial {C_2}^2}\frac{\partial F_0}{\partial C_1} - \frac{\partial F_1}{\partial C_2}\frac{\partial^2 F_0}{\partial C_1 \partial C_2}\right]\frac{\partial F_0}{\partial C_1} \\ &- \left[\frac{\partial^2 F_1}{\partial {C_1}^2}\frac{\partial F_0}{\partial C_2} + \frac{\partial F_1}{\partial C_1}\frac{\partial^2 F_0}{\partial C_1 \partial C_2} - \frac{\partial^2 F_1}{\partial C_1 \partial C_2}\frac{\partial F_0}{\partial C_1} - \frac{\partial F_1}{\partial C_2}\frac{\partial^2 F_0}{\partial {C_1}^2}\right]\frac{\partial F_0}{\partial C_2} \neq 0\end{aligned} \tag{13}$$

then as sample size $n \to \infty$, $\sqrt{n}\,(\widehat{\text{sen}}_C - \text{sen}_C)$ converges to a normal distribution with mean 0 and variance given by (19).

Remark 1. *Condition (13) can be justified algebraically for bivariate normal random variables when F_1 and F_0 have a different covariance matrix.*

Remark 2. *Although the asymptotic normality holds for all the three estimators under fairly mild conditions, the asymptotic variances of the sensitivities are hard to estimate directly. Therefore for the inference, their standard errors are estimated using the nonparametric bootstrap method (Efron and Tibshirani, 1994). Specifically, 200 samples with the same size are drawn from the original data with replacement. For each of the three diagnostic tests, each sample yields an estimated sensitivity at the given specificity, and the standard error is then estimated by the standard deviation of the 200 estimated sensitivities.*

5 Numeric Results

5.1 Application: ELISA Tests for E2 antibodies

In this section, we applied all three methods to the data set from the E2 antibody study example described in the introduction with the goal of detection of the antibody presence in the blood sample of 100 HIV infected study participants. The scatter plots of Figure 3 presents the results from the two ELISA tests. The data are fitted by the mixture of two bivariate normal distributions: $N(\mu_1, V_1)$ and $N(\mu_0, V_0)$. The maximum likelihood estimation gives $\hat{\mu}_1 = (1.01, 0.84)^T$, $\hat{\mu}_0 = (0.16, 0.24)^T$, and

$$\hat{V}_1 = \begin{pmatrix} 0.54 & 0.22 \\ 0.22 & 0.40 \end{pmatrix}, \hat{V}_0 = \begin{pmatrix} 0.004 & 0.001 \\ 0.001 & 0.017 \end{pmatrix}.$$

The estimated ROC curves based on the two individual tests are depicted in Figure 4. It appears that Test 1 is superior to Test 2 as it is more sensitive to the antibody presence for any given specificity and hence it is chosen to be the initial test for our proposed sequential method.

The decision rules for the three optimal composite tests with specificity $p_0 = 0.90$ are superimposed to the scatter plot of Figure 3 to characterise the composite tests. Their ROC curves are also plotted in Figure 4 along with those based on the individual tests. Figure 4 indicates that all the optimal composite tests provide a better discriminant capability than the two individual tests for

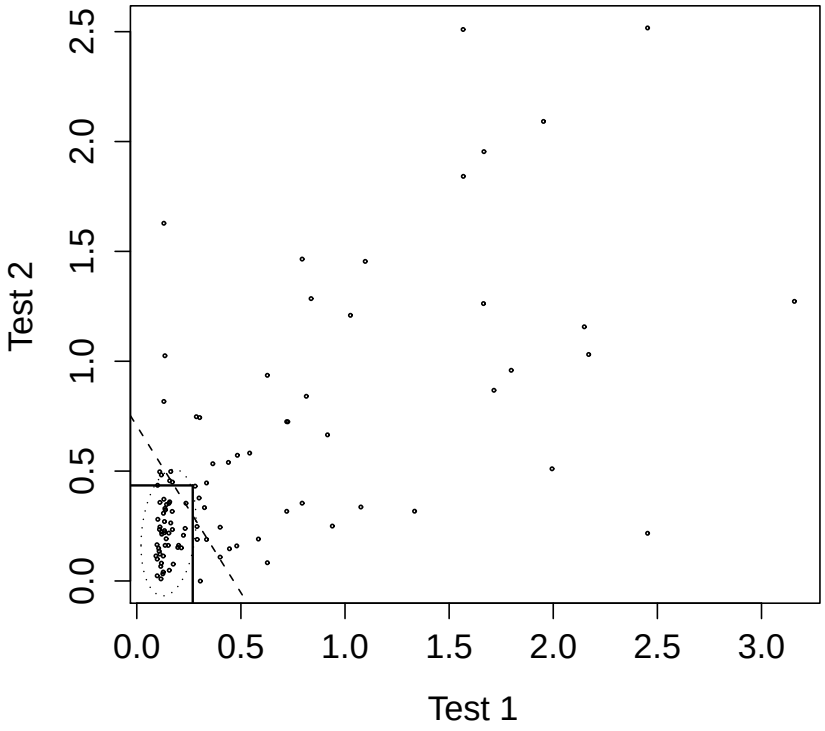

Figure 3: Results from the two tests in 100 blood samples along with the optimal linear composite test (dashed line), the optimal risk score composite test (dotted line) and the optimal sequential composite test (solid line) at specificity =0.90.

this application. Both the optimal sequential and optimal risk score composite tests are substantially better than the individual tests, but the optimal linear composite test only adds little discriminant power to the best individual test. While the optimal risk score composite test is superior to the optimal sequential composite test as anticipated due to Neyman-Pearson theorem of likelihood ratio test, the optimal sequential composite test only needs 47% of the second tests for the blood samples.

5.2 Simulation Studies

In this section, we assess the performance of the three methods using the simulation studies described below.

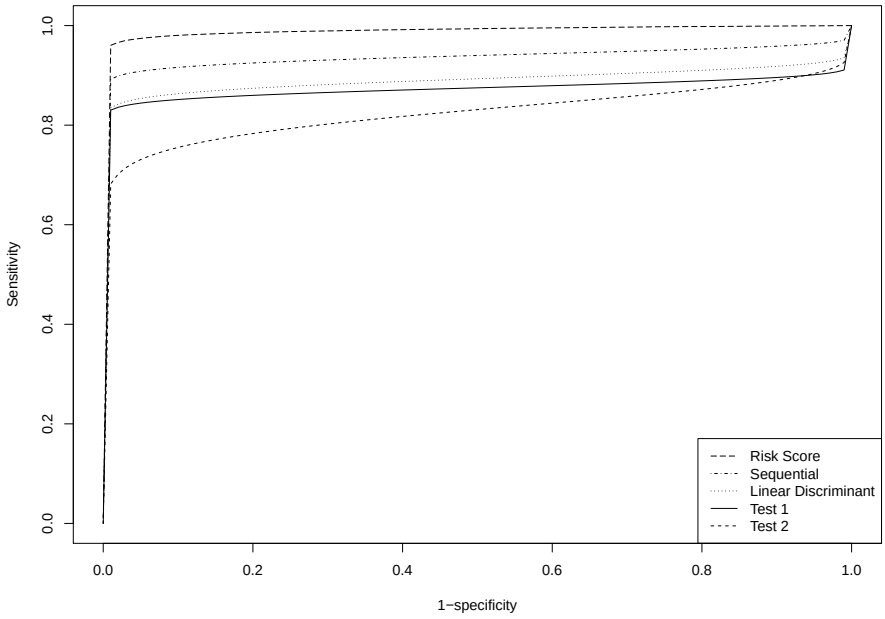

Figure 4: ROC curves for Test 2, Test 1, the optimal linear, sequential and risk score composite tests (from bottom to top).

In the first study, we generate two diagnostic markers for the case group from a bivariate normal distribution $N(\mu_1, V_1)$ of

$$\mu_1 = (3.77, 1.51)^T \text{ and } V_1 = \begin{pmatrix} 3.97 & 0.69 \\ 0.69 & 1.42 \end{pmatrix},$$

and the markers for the non-case group from a bivariate normal distribution $N(\mu_0, V_0)$ of

$$\mu_0 = (2, 0.81)^T \text{ and } V_0 = \begin{pmatrix} 0.68 & 0.03 \\ 0.03 & 0.18 \end{pmatrix}.$$

Because it is important to evaluate the effects of sample size and latent case prevalence on the performance of the methods, three total sample sizes (100, 200 and 400) with case prevalence of 0.25 and 0.5, respectively, are examined. For each combination of the sample size and case prevalence, 1000 Monte-Carlo samples are generated from the two designed bivariate normal distributions, $N(\mu_1, V_1)$ and $N(\mu_0, V_0)$. For inference, the standard error of the estimated sensitivity is estimated via the nonparametric bootstrap method aforementioned and its 95% Wald confidence interval is constructed using the bootstrap standard error as well.

We note that the exact sensitivity at a given specificity can be analytically determined for the composite tests for the bivariate normal distributions F_1 and F_0 for the optimal linear composite test and the optimal sequential test. The exact sensitivity for the optimal risk score test could be approximated by the aforementioned Monte-Carlo method. Therefore, for each simulated sample, we can estimated model-based specificity and sensitivity with F_1 and F_0 replaced by their corresponding ML estimates of $\hat{F}_1$ and $\hat{F}_0$ and hence the bias, root mean square error (RMLE) and coverage probability of the 95% confidence interval can be estimated with the Monte-Carlo simulation study. In addition, since the true case status is known for the simulated data, the empirical specificity and sensitivity can be also directly computed for the three composite tests. Table 1 summarises this simulation study based on the 1000 Monte-Carlo samples. As seen in the table, the composite tests are generally better than the best individual test (Test 1) with the optimal risk score composite test having the largest sensitivity for a given specificity. Though slightly less sensitive than the optimal risk score composite test, the optimal sequential composite test has a better sensitivity than the optimal linear composite test. The model-based estimated sensitivities appear to be unbiased with decreasing RMSE and right coverage probability when sample size increases, which justifies the asymptotic normality properties stated in Section 4. By comparing the RMLE, it can be also inferred that the estimated sensitivities may be more accurate when the case prevalence increases. Moreover, the empirical specificity and sensitivity for the composite tests are all in-line with their designed values for the simulation study, indicating these tests working properly.

The second simulation study is designed to evaluate the robustness of the tests against the normality assumption. Since the multivariate normal distribution is a special form of the Gaussian copula (Nelsen, 1999), with marginally

Table 1: The bias (Bias) and the root mean square error (RMSE) of the estimated sensitivity, the average of the empirical sensitivity (ESen) and the empirical specificity (ESpe), and the coverage probability of 95% confidence intervals (CP) at the given specificities based on 1000 Monte-Carlo samples. N is the total sample size in two groups. π is the case prevalence of the samples.

$\pi = 0.25$										
Test 1 alone										
	Specificity=80%, Sensitivity=0.705					Specificity=90%, Sensitivity=0.640				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.038	0.145	0.701	0.797	0.836	0.044	0.138	0.634	0.898	0.852
200	0.030	0.113	0.710	0.799	0.874	0.032	0.099	0.641	0.900	0.900
400	0.010	0.077	0.705	0.799	0.945	0.011	0.070	0.640	0.899	0.948
Optimal Linear Composite Test										
	Specificity=80%, Sensitivity=0.738					Specificity=90%, Sensitivity=0.682				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.058	0.149	0.742	0.793	0.741	0.064	0.141	0.685	0.892	0.771
200	0.039	0.113	0.743	0.797	0.844	0.042	0.100	0.686	0.897	0.867
400	0.014	0.074	0.739	0.798	0.942	0.014	0.066	0.684	0.898	0.946
Optimal Sequential Composite Test										
	Specificity=80%, Sensitivity=0.802					Specificity=90%, Sensitivity=0.750				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.037	0.119	0.794	0.794	0.777	0.044	0.116	0.741	0.895	0.822
200	0.026	0.092	0.800	0.798	0.856	0.029	0.084	0.747	0.899	0.872
400	0.009	0.063	0.800	0.798	0.941	0.010	0.059	0.749	0.899	0.947
Optimal Risk Score Composite Test										
	Specificity=80%, Sensitivity=0.864					Specificity=90%, Sensitivity=0.809				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.033	0.087	0.843	0.790	0.765	0.045	0.080	0.787	0.891	0.781
200	0.020	0.068	0.853	0.798	0.867	0.027	0.060	0.798	0.899	0.878
400	0.007	0.050	0.861	0.799	0.943	0.011	0.045	0.807	0.898	0.945
$\pi = 0.5$										
Test 1 alone										
	Specificity=80%, Sensitivity=0.705					Specificity=90%, Sensitivity=0.640				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.031	0.120	0.698	0.800	0.882	0.034	0.111	0.630	0.896	0.903
200	0.018	0.084	0.708	0.799	0.929	0.019	0.072	0.642	0.899	0.939
400	0.006	0.065	0.704	0.800	0.957	0.007	0.060	0.637	0.900	0.961
Optimal Linear Composite Test										
	Specificity=80%, Sensitivity=0.738					Specificity=90%, Sensitivity=0.682				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.044	0.122	0.732	0.792	0.818	0.048	0.111	0.674	0.891	0.839
200	0.023	0.081	0.741	0.795	0.917	0.024	0.067	0.684	0.896	0.932
400	0.009	0.061	0.737	0.799	0.959	0.010	0.054	0.681	0.898	0.960
Optimal Sequential Composite Test										
	Specificity=80%, Sensitivity=0.802					Specificity=90%, Sensitivity=0.750				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.029	0.098	0.788	0.795	0.848	0.034	0.091	0.730	0.893	0.894
200	0.016	0.066	0.800	0.796	0.927	0.018	0.056	0.747	0.896	0.951
400	0.007	0.053	0.800	0.799	0.947	0.007	0.048	0.747	0.898	0.961
Optimal Risk Score Composite Test										
	Specificity=80%, Sensitivity=0.863					Specificity=90%, Sensitivity=0.811				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.031	0.076	0.836	0.786	0.765	0.039	0.064	0.780	0.881	0.788
200	0.014	0.054	0.855	0.794	0.908	0.017	0.043	0.800	0.891	0.920
400	0.007	0.045	0.859	0.798	0.932	0.007	0.040	0.805	0.896	0.941

normal distributed random variables, we generate the data from a mixture of two Gaussian copulas with the same correlation in each group as in the previous simulation study. However, the marginal distributions of the two markers are set to be Student-t with 4 degrees of freedom and scaled to have the same means and variances as in the previous simulation study. Although the data have the same means and variances as in the first study, the distribution of the diagnostic markers are misspecified from the mixture of bivariate normal distribution. Figure 5 displays one data set of 100 subjects with the case prevalence 0.5 that also appears similar to the situation presented in the motivating example.

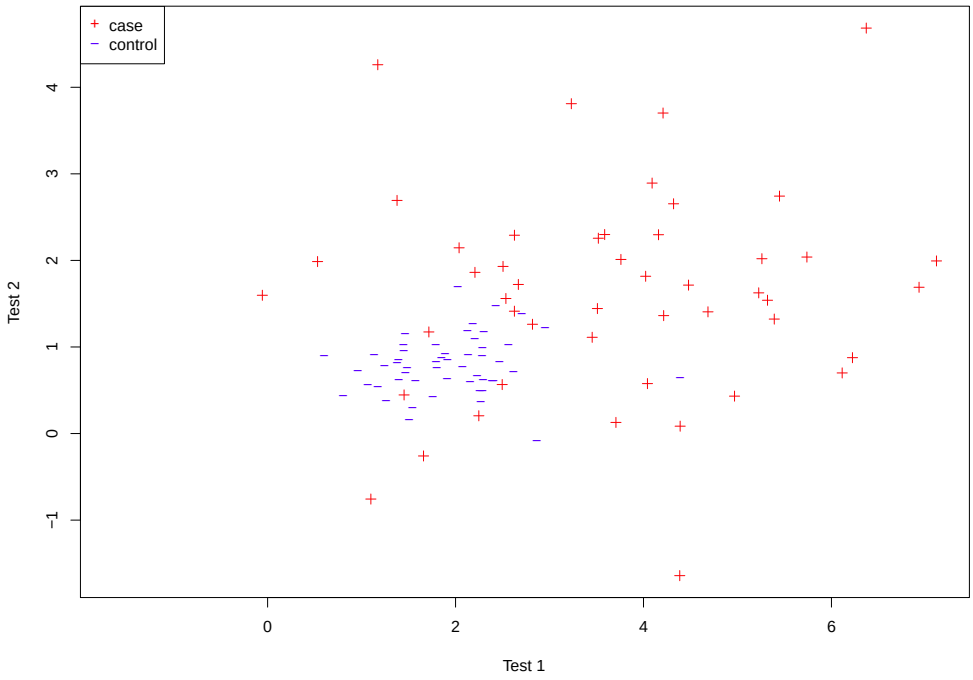

Figure 5: Scatter plot of a simulated data set of 100 subjects from the Gaussian copula model with the Student-t marginal. The case prevalence is 0.5.

The data are still fitted using the mixture of two bivariate normal distributions and the parallel results are summarised in Table 2. We note that, except for the optimal linear composite, the exact sensitivities for the individual test (Test 1), the optimal risk score and optimal sequential composite tests are still able to be analytically determined based on the true Student-t distributions, F_0 and F_1. It is interesting to observe that although the (normal) model-based estimated sensitivity from Test 1 alone is seemingly biased from the designed value, the biases of both the optimal risk score and sequential composite methods tend

to be negligible. The empirical specificity and sensitivity of the composite tests are also close to their corresponding designed values indicating the composite tests are fairly robust in terms of accurately classifying the study subjects even though the underlying statistical model for developing the tests are misspecified. While the estimated sensitivities based on the composite tests appear consistent, the asymptotic normality of the estimates are no longer valid as indicated by the coverage probability of the confidence interval. This is not surprising, because the asymptotic normality of the estimated model parameters depends on the assumption that the underlying statistical model is correctly specified. Therefore, it should be cautious in making distributional inference about the estimated sensitivity.

Table 2: The bias (Bias) and the root mean square error of the estimated sensitivity (RMSE), the average of the empirical sensitivity (ESen), the average of the empirical specificity (ESpe), and the coverage of 95% confidence intervals (CP) at the given specificities based on 1000 Monte-Carlo samples of Gaussian copula with the Student-t marginal distributions . N is the total sample size in two groups. π is the case prevalence of the samples.

$\pi = 0.25$										
Test 1 alone										
	Specificity=80%, Sensitivity=0.783					Specificity=90%, Sensitivity=0.716				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.106	0.134	0.775	0.806	0.809	-0.092	0.170	0.717	0.890	0.854
200	-0.123	0.118	0.781	0.808	0.586	-0.110	0.163	0.724	0.893	0.663
400	-0.131	0.113	0.778	0.810	0.200	-0.118	0.161	0.720	0.893	0.295
Optimal Linear Composite Test										
	Specificity=80%, Sensitivity=NA†					Specificity=90%, Sensitivity=NA†				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	NA	NA	0.790	0.790	NA	NA	NA	0.742	0.877	NA
200	NA	NA	0.796	0.796	NA	NA	NA	0.747	0.884	NA
400	NA	NA	0.794	0.798	NA	NA	NA	0.747	0.886	NA
Optimal Sequential Composite Method										
	Specificity=80%, Sensitivity=0.856					Specificity=90%, Sensitivity=0.786				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.061	0.094	0.853	0.780	0.894	-0.034	0.119	0.807	0.864	0.947
200	-0.079	0.077	0.859	0.784	0.714	-0.052	0.110	0.812	0.868	0.940
400	-0.085	0.070	0.858	0.786	0.321	-0.059	0.106	0.814	0.868	0.744
Optimal Risk Score Composite Test										
	Specificity=80%, Sensitivity=0.896					Specificity=90%, Sensitivity=0.830				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.008	0.064	0.898	0.739	0.896	0.034	0.058	0.850	0.830	0.807
200	-0.002	0.052	0.904	0.739	0.953	0.022	0.049	0.859	0.829	0.860
400	-0.004	0.047	0.907	0.739	0.944	0.020	0.043	0.866	0.827	0.815
$\pi = 0.5$										
Test 1 alone										
	Specificity=80%, Sensitivity=0.783					Specificity=90%, Sensitivity=0.716				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.092	0.121	0.739	0.835	0.902	-0.081	0.159	0.668	0.909	0.942
200	-0.099	0.107	0.746	0.838	0.727	-0.088	0.147	0.678	0.910	0.844
400	-0.105	0.093	0.751	0.839	0.347	-0.093	0.141	0.684	0.911	0.564
Optimal Linear Composite Test										
	Specificity=80%, Sensitivity=NA†					Specificity=90%, Sensitivity=NA†				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	NA	NA	0.757	0.824	NA	NA	NA	0.697	0.901	NA
200	NA	NA	0.762	0.828	NA	NA	NA	0.705	0.903	NA
400	NA	NA	0.770	0.832	NA	NA	NA	0.715	0.907	NA
Optimal Sequential Composite Method										
	Specificity=80%, Sensitivity=0.856					Specificity=90%, Sensitivity=0.786				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.049	0.085	0.819	0.811	0.952	-0.022	0.106	0.758	0.885	0.961
200	-0.056	0.071	0.826	0.810	0.899	-0.029	0.095	0.770	0.884	0.986
400	-0.059	0.056	0.836	0.813	0.612	-0.032	0.086	0.782	0.886	0.968
Optimal Risk Score Composite Test										
	Specificity=80%, Sensitivity=0.895					Specificity=90%, Sensitivity=0.831				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.014	0.067	0.859	0.756	0.822	0.041	0.057	0.802	0.837	0.704
200	0.008	0.057	0.868	0.757	0.912	0.033	0.047	0.816	0.837	0.773
400	0.004	0.052	0.883	0.757	0.920	0.028	0.041	0.834	0.839	0.740

† The best linear combination cannot be calculated directly from the true model because the joint distribution of two tests in each group is not multivariate normal.

Table 3: Average of percentage of subjects taking only one test in simulation studies (%) under different sample sizes, case prevalences, and pre-specified specificities.

	π =0.25				π =0.5			
	Spe=80%		Spe=90%		Spe=80%		Spe=90%	
N	Normal	t	Normal	t	Normal	t	Normal	t
100	28.1	29.6	20.9	23.4	40.7	41.6	33.9	35.6
200	27.5	28.8	20.4	23.0	40.6	41.2	33.9	35.4
400	27.1	28.3	20.2	22.7	40.2	41.2	33.5	35.5

In the simulation studies, although it is less accurate in classification than the optimal risk score composite test, the optimal sequential composite test does not require all subjects to take both tests. As listed in Table 3, when the marginal distribution of the two assays are normal, using the optimal sequential rule, around 27% of subjects in the sample only needs to take one test to obtain the diagnosis of the case, which prevalence is 25%. This percentage is higher, around 40%, for the case with a prevalence of 50%. The percentages are similar for the data with student-t marginal distributions since the patterns of the data are similar.

6 Final Remarks

In this paper, we have overviewed and extended existing methods for combining multiple quantitative tests to classify subjects without a "gold standard". In addition, we also implemented an alternative classification method based on sequential tests under no "gold standard" circumstance. All the extended methods are developed under the model assumption that the diagnostic markers for the tests for both case and non-case populations come from multivariate normal distributions. A mixture model of two multivariate normal distributions is fitted to the unclassified data and the optimal decision rule for each method is determined with the fitted model. All the methods are illustrated with the real data set for two ELISA tests to detect GBV-C E2 antibody in participating subjects, and their performance is assessed through simulation studies. For the data presented in Figure 3 and Figure 5, the simulation studies demonstrate that the optimal risk score composite test has the most discriminant power to distinguish case from non-case in view of AUC and is most sensitive among the three composite tests. The optimal sequential composite test, though little less sensitive than its risk score counterpart, outperforms the optimal linear composite test and has an additional advantage of engaging less tests. This advantage of the sequential method is particularly desired when the tests are costly or not applicable to all study subjects in some applications. The optimal classification rules of the tests in this article is purely based on the classification accuracy without considering risk or cost associated with the tests. If the risk or cost ought to be considered for determining an optimal decision rule in some

applications, the risk score composite test may not be the optimal test among the three.

The distributional assumption plays a key role throughout this paper. First, the optimal linear composite test is specifically designed for multivariate normal markers only. Second, the model based inference is valid only when the underlying statistical model is indeed multivariate normal. For the optimal risk score composite test, the underlying model assumption can be relaxed if there is "gold standard" resulting in correct classification. McIntosh and Pepe (2002) used logistic regression model to develop the optimal risk score composite test. However if the logistic regression model is misspecified, the test performance may not be optimized. In our extra simulation study (not shown here) for multivariate normal diagnostic markers, the optimal risk score composite test with the MnIntosh-Pepe's logistic regression model would largely reduce the potential sensitivity for normal data but the normality distribution is not utilized. In practice, it is hard to examine the normality assumption on the unclassified data set and the "working" model of multivariate normal distribution is probably the most popular option of parametric approach. Not only does it provide an easily implemented estimation procedure, it also results in a quite robust classification rule against the underlying statistical models as long as the markers possibly resemble normal data, which is exemplified in our second simulation study.

When multivariate normal distribution cannot be justified from a biological perspective and is really questionable for the unclassified data in applications, one may consider possible monotone transformations to convert the original data to normal-like multivariate data and apply the methods to the transformed data. The monotone transformation will likely result in same classification as the methods would do for the original data. Alternatively, one may consider nonparametric approaches. Hall and Zhou (2003) proposed a mixture nonparametric model to unclassified data, but their method is very complicated to implement and is restrictive. The tensor spline-based sieve maximum likelihood estimation (Wu, 2010) of the multivariate distribution function is a compromise to the Hall-Zhou's nonparametric estimation of mixture distribution. Although the optimal risk score and sequential composite tests can still be computed with the tensor spline-based sieve estimation, the numerical implementation of the tests is much more demanding and challenging than the multivariate normal model. Moreover, the statistical properties of the tests using the spline-based model are much harder to study. This also remains as an open problem for further investigation.

There are multiple ways to design a sequential composite test according to the data that reflect the study nature. The sequential test designed in this paper is evidently supported by the biological mechanism in E2 antibody. For a blood sample, it is known that each ELISA test returns a value reflecting the concentration of the E2 antibody and a larger value implies a greater chance of the antibody presence. Because antibodies have different binding sites on the viral protein, the binding site of the E2 antibody detected in the ELISA may not be recognised by some individuals, or the methodology may render the binding site(s) of other human antibodies inaccessible in the assay. In this

circumstance, the variance of the ELISA results in GBV-C E2 antibodies present (case) population is greater than the variances in GBV-C E2 antibodies negative (non-case) population. Graphically, a cluster of small values of the ELISA tests represents the GBV-C E2 antibodies negative population. So the sequential method starting with classification for case is the best option. If, for some applications, there is a small cluster of points present in the upper right corner of the scatter plot that likely represents the group of case, it would be more reasonable to start with classification of non-case as shown in Figure 6 for the sequential test.

The optimal sequential composite test is statistically equivalent to the implementation of a sequence of tests discussed by Thompson (2003). In that paper, she focused on evaluating a sequence of tests, and only provided the theoretical development. The choice of the optimal threshold and statistical inferences on the estimated ROC curves were not addressed. Besides, her results are based on an available "gold standard". But the optimal sequential composite test developed here has a limitation in generalization to more than two individual tests. Although, the proposed sequential test can be similarly designed for the situation with more than two tests, it is, however, a mathematically challenging problem because finding the optimal cut-off values cannot be equivalently converted to the problem of solving a nonlinear system as it does for the two-test case. The grid search is a straightforward option but it can be very numerically inefficient, especially for high dimensional data.

Appendix A Proof of Theorem 1

Throughout the rest of the paper, K is denoted as a universal constant that may vary from place to place. For each n, $X_{n,1}, X_{n,2}, \ldots, X_{n,m(n)}$, where $m(n)/n \to \infty$ as $n \to \infty$, are *i.i.d.* random variables according to a probablility measure P_n, which converges to a measure P_0 in a suitable sense. Let $\mathcal{F}$ be the collection of all indicator functions of form $f_t(x) = \mathrm{I}_{[x \le t]}$, with t ranging over $\mathbb{R}$. Define the semimetric ρ_P on $\mathcal{F}$ as

$$\rho_P(f) = \left(P|f|^2\right)^{1/2}, \forall f \in \mathcal{F},$$

and define $l^\infty(\mathcal{F})$ as the set of all uniformly bounded functionals on $\mathcal{F}$: $z : \mathcal{F} \mapsto \mathbb{R}$ such that

$$\sup_{f \in \mathcal{F}} |z(f)| < \infty.$$

$\mathbb{P}_n$ is the ordinary empirical measure based on the sample of $X_{n,1}, \cdots, X_{n,m(n)}$, that is

$$\mathbb{P}_n = \sum_{i=1}^{m(n)} \delta_{X_{n,i}},$$

and the centered empirical measure of $\mathbb{P}_n$ is defined as

$$\mathbb{G}_{n,P_n} = \sqrt{m(n)}(\mathbb{P}_n - P_n).$$

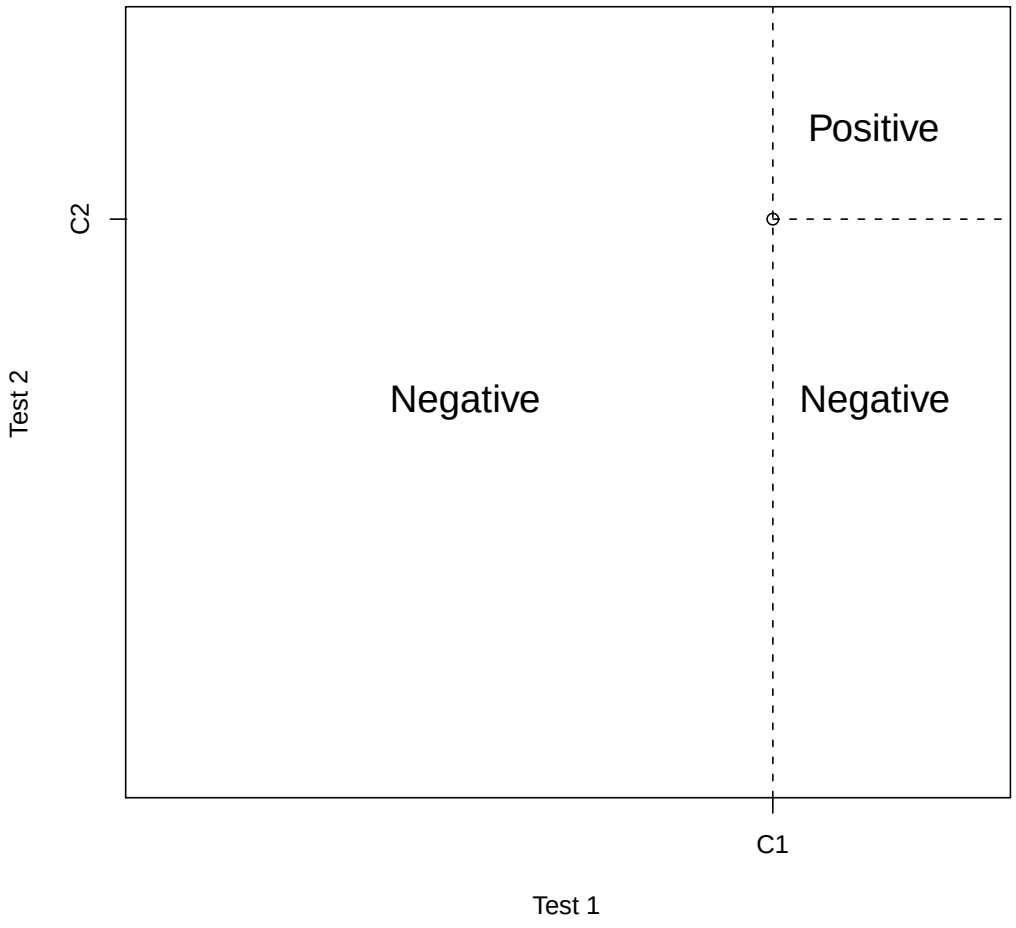

Figure 6: Illustration of the alternative 2-cutoff sequential classification method.

(van der Vaart and Wellner, 1996, Ch. 2)

The following lemmas are utilized to prove Theorem 1.

Lemma 1. *If the semimetric ρ_{P_n} converges uniformly to ρ_{P_0} in the sense that*

$$\sup_{f,g\in\mathcal{F}} |\rho_{P_n}(f,g) - \rho_{P_0}(f,g)| \to 0, \quad as \quad n \to \infty \tag{14}$$

then $\mathbb{G}_{n,P_n}$ converges in distribution to $\mathbb{G}_{P_0}$ in $l^\infty(\mathcal{F})$.

Proof. We prove this lemma using Theorem 2.8.10 of van der Vaart and Wellner (1996). $\mathcal{F}$ contains all the indicator functions of form $f_t(x) = \mathrm{I}_{[x\le t]}$ for $t \in \mathbb{R}$, so it is a class of measurable functions. Since the indicator function takes values at only 0 and 1, the constant function $F = 1$ is an envelope function for $\mathcal{F}$, which is measurable and totally bounded for ρ_{P_0}. Moreover, $\forall\epsilon > 0$, $\exists N$, such that $\forall n \ge N$, $\epsilon\sqrt{m(n)} > 1$ and $F^2\left\{F \ge \epsilon\sqrt{m(n)}\right\} = 0$. Hence, the condition that $\limsup_{n\to\infty} P_n F^2\left\{F \ge \epsilon\sqrt{m(n)}\right\} = 0$ is satisfied.

Next, we show that the class $\mathcal{F}$ is P_n-Donsker. For each n and for any $\epsilon > 0$, assuming that H_n is the distribution function induced by the probability measure P_n, we construct the brackets of the form $[\mathrm{I}_{(-\infty,t_{i-1}]}, \mathrm{I}_{(-\infty,t_i]}]$ with a grid of points $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ satisfying $H_n(t_i-) - H_n(t_{i-1}) < \epsilon$ for each i. This can be achieved by the fact that P_n is a probability measure that converges to P_0. These brackets have $L_1(P_n)$-size ϵ, and the total number k is bounded by $1/\epsilon$. Because $P_n f^2 \le P_n f$ for every $0 \le f \le 1$, the $L_2(P_n)$-size of the brackets is bounded by $\sqrt{\epsilon}$. Thus we have the bracketing number $N_{[\,]}(\sqrt{\epsilon}, \mathcal{F}, L_2(P_n)) \le (1/\epsilon)$. Equivalently, $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n)) \le (1/\epsilon^2)$. The bracketing entropy of $\mathcal{F}$ is of the order of $\log(\epsilon)$, which is $o(1/\epsilon^\tau)$ for any $\tau \in (0,1)$ since $\lim_{\epsilon\to 0} \epsilon^\tau \log(\epsilon) = 0$. Therefore, the bracketing integral

$$\begin{aligned} J_{[\,]}(\delta, \mathcal{F}, L_2(P_n)) &= \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n))}d\epsilon \\ &\le K\int_0^\delta \epsilon^{-\tau/2}d\epsilon = \frac{K}{1-\tau/2}\delta^{1-\tau/2} \\ &< \infty, \text{ for any } \delta > 0. \end{aligned}$$

This indicates that $\mathcal{F}$ is a P_n-Donsker class, and it also indicates that

$$\lim_{\delta_n\to 0} J_{[\,]}(\delta_n, \mathcal{F}, L_2(P_n)) = 0.$$

Therefore, all the conditions in Theorem 2.8.10 of van der Vaart and Wellner (1996) are satisfied, so $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$ in $l^\infty(\mathcal{F})$.

□

Lemma 2. *Suppose that the condition* (14) *in Lemma 1 is satisfied. If f_n is a sequence of functions in $\mathcal{F}$ such that $\int (f_n - f_0)^2 dP_0$ converges to 0 in probability for some $f_0 \in \mathcal{F}$, then $\mathbb{G}_{n,P_n}(f_n - f_0) \to_P 0$ and $\mathbb{G}_{n,P_n} f_n \rightsquigarrow \mathbb{G}_{P_0} f_0$.*

Proof. The proof of this lemma is similar to the proof of Lemma 19.24 of van der Vaart (2000). Define a functional $g : l^\infty(\mathcal{F}) \times \mathcal{F} \mapsto \mathbb{R}$ by $g(z, f) = z(f) - z(f_0)$. The semimetrics on $l^\infty(\mathcal{F})$ and $\mathcal{F}$ are denoted by $\|\cdot\|_\infty$ and $\|\cdot\|_2$, respectively. We need to show that g is continuous with respect to the product semimetric at every (z, f) such that $f \mapsto z(f)$ is continuous.

Actually, if $(z_n, f_n) \to (z, f)$ in $l^\infty(\mathcal{F}) \times \mathcal{F}$, then $\|z_n - z\|_\infty \to 0$, and $\|f_n - f\|_2 \to 0$. Hence,

$$\begin{aligned}
|g(z_n, f_n) - g(z, f)| &= |\,(z_n(f_n) - z_n(f_0)) - (z(f) - z(f_0))\,| \\
&= |z_n(f_n) - z(f_n) + z(f_n) - z(f) - (z_n(f_0) - z(f_0))\,| \\
&\le |z_n(f_n) - z(f_n)| + |z(f_n) - z(f)| + |z_n(f_0) - z(f_0)| \\
&\le \|z_n - z\|_\infty + |z(f_n) - z(f)| + \|z_n - z\|_\infty \\
&= o(1) \text{ if } z \text{ is continuous at } f.
\end{aligned}$$

Now let $z_n = \mathbb{G}_{n,P_n}$, and $z = \mathbb{G}_{P_0}$. By the assumption, $f_n \to_P f_0$ in metric space $\mathcal{F}$, and by Lemma 1, $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$ in $l^\infty(\mathcal{F})$. Then $(\mathbb{G}_{n,P_n}, f_n) \rightsquigarrow (\mathbb{G}_{P_0}, f_0)$ in the space $l^\infty(\mathcal{F}) \times \mathcal{F}$. By Lemma 18.15 of van der Vaart (2000), $\mathbb{G}_{P_0}$ is continuous on $\mathcal{F}$ for almost all sample paths, thus g is continuous at almost every point of $(\mathbb{G}_{P_0}, f_0)$. By the continuous mapping theorem, $\mathbb{G}_{n,P_n}(f_n - f_0) = g(\mathbb{G}_{n,P_n}, f_n) \rightsquigarrow g(\mathbb{G}_{P_0}, f_0) = 0$. It is equivalent to $\mathbb{G}_{n,P_n}(f_n - f_0) \to_P 0$.

The second assertion follows because $\mathbb{G}_{n,P_n} f_n = o_P(1) + \mathbb{G}_{n,P_n} f_0 \rightsquigarrow \mathbb{G}_{P_0} f_0$ due to the result of Lemma 1. □

Lemma 3. *Suppose F_n is the underlying distribution function of $X_{n,1}, \ldots, X_{n,m(n)}$ with $F_n \rightsquigarrow F_0$[1], and $\mathbb{F}_n$ is the empirical distribution function of $X_{n,1}, \ldots, X_{n,m(n)}$. If F_n and F_0 are continuously differentiable at $F_n^{-1}(c)$ and $F_0^{-1}(c)$ with bounded and strictly positive derivative $f_n\left(F_n^{-1}(c)\right)$ and $f_0\left(F_0^{-1}(c)\right)$ for $\forall c \in (0, 1)$, respectively, then, $\sqrt{m(n)}\left(\mathbb{F}_n^{-1}(c) - F_n^{-1}(c)\right)$ converges in distribution to a normal distribution with mean 0 and variance $c(1-c)/f_0^2\left(F_0^{-1}(c)\right)$.*

Proof. Define a function ϕ as $\phi(P) = F^{-1}(c)$, then $\phi(P_n) = F_n^{-1}(c)$ and $\phi(\mathbb{P}_n) = \mathbb{F}_n^{-1}(c)$. By the *von Mises expansion* (van der Vaart, 2000, Ch. 20),

$$\phi(\mathbb{P}_n) - \phi(P_n) \approx \frac{1}{\sqrt{m(n)}}\phi'(\mathbb{G}_{n,P_n}) = \frac{1}{m(n)}\sum_{i=1}^{m(n)} \phi'_{P_n}(\delta_{X_{n,i}} - P_n),$$

where the influence function $\phi'_{P_n}(\delta_x - P_n)$ can be computed as

$$\phi'_{P_n}(\delta_x - P_n) = \frac{d}{dt}_{|t=0} \phi\left((1-t)P_n + t\delta_x\right).$$

Let $I_n(x) = \phi'_{P_n}(\delta_x - P_n)$ and $P_{ntx} = (1-t)P_n + t\delta_x$. For any t, $F_{P_{ntx}}\left(\phi(P_{ntx})\right) = (1-t)F_n\left(\phi(P_{ntx})\right) + tG_x\left(\phi(P_{ntx})\right) = c$, where $G_x(t) = \mathrm{I}_{[x,\infty)}(t)$. Taking the

[1] $F_n \rightsquigarrow F_0$ if and only if $F_n(t) \to F_0(t)$ at every t where F_0 is continuous.

derivative with respect to t and evaluating it at $t = 0$ on both sides, we get

$$\begin{aligned} &-F_n\left(\phi(P_n)\right) + f_n\left(\phi(P_n)\right) I_n(x) + G_x\left(\phi(P_n)\right) \\ &\qquad = -c + f_n\left(\phi(P_n)\right) I_n(x) + \mathrm{I}_{[x \le \phi(P_n)]} = 0. \end{aligned}$$

So the influence function is given by

$$I_n(x) = -\frac{\mathrm{I}_{[x \le \phi(P_n)]} - c}{f_n\left(\phi(P_n)\right)},$$

and

$$\begin{aligned} \sqrt{m(n)}\left(\mathbb{F}_n^{-1}(c) - F_n^{-1}(c)\right) &= \sqrt{m(n)}\left(\phi(\mathbb{P}_n) - \phi(P_n)\right) = \frac{1}{\sqrt{m(n)}} I_n(X_{n,i}) + o_P(1) \\ &= \frac{1}{f_n\left(F_n^{-1}(c)\right)} \cdot \frac{1}{\sqrt{m(n)}} \sum_{i=1}^{m(n)} \left(\mathrm{I}_{[X_{n,i} \le F_n^{-1}(c)]} - c\right) + o_P(1). \end{aligned}$$

Note that $f_n = F_n'$ and $f_0 = F_0'$ are the density functions of F_n and F_0, respectively. $F_n^{-1}(c) \to F_0^{-1}(c)$, because $F_n \rightsquigarrow F_0$ by Lemma 21.2 of van der Vaart (2000). Hence $f_n\left(F_n^{-1}(c)\right) \to f_0\left(F_0^{-1}(c)\right)$.

Furthermore, $\mathrm{I}_{(-\infty, F_n^{-1}(c)]}$ is a sequence of functions in $\mathcal{F}$ that converges to $\mathrm{I}_{(-\infty, F_0^{-1}(c)]}$ in the sense that

$$\int \left(\mathrm{I}_{(-\infty, F_n^{-1}(c)]} - \mathrm{I}_{[(-\infty, F_0^{-1}(c)]}\right)^2 dP_0 = \left| \int_{F_0^{-1}(c)}^{F_n^{-1}(c)} dP_0 \right| \to 0.$$

Thus by Lemma 2,

$$\frac{1}{\sqrt{m(n)}} \sum_{i=1}^{m(n)} \left(\mathrm{I}_{[X_{n,i} \le F_n^{-1}(c)]} - c\right) = \mathbb{G}_{n,P_n} \mathrm{I}_{[(-\infty, F_n^{-1}(c)]} \rightsquigarrow \mathbb{G}_{P_0} \mathrm{I}_{[(-\infty, F_0^{-1}(c)]}.$$

Taking together, $\sqrt{m(n)}\left(\mathbb{F}_n^{-1}(c) - F_n^{-1}(c)\right) \rightsquigarrow \frac{1}{f_0^2\left(F_0^{-1}(c)\right)} \mathbb{G}_{P_0} \mathrm{I}_{[(-\infty, F_0^{-1}(c)]}$, which is the normal distribution with mean 0 and variance $c(1-c)/f_0^2\left(F_0^{-1}(c)\right)$.

□

Now using the preceding lemmas, Theorem 1 is proved as follows.
It is noted that

$$\begin{aligned} &\sqrt{n}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}_{0,\hat{\theta}_n,m}^{-1}(q_0)\right) - H_{1,\theta}\left(H_{0,\theta}^{-1}(q_0)\right)\right] \\ &= \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}_{0,\hat{\theta}_n,m}^{-1}(q_0)\right) - H_{1,\hat{\theta}_n}\left(\mathbb{H}_{0,\hat{\theta}_n,m}^{-1}(q_0)\right)\right] \qquad (15) \\ &+ \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}\left[H_{1,\hat{\theta}_n}\left(\mathbb{H}_{0,\hat{\theta}_n,m}^{-1}(q_0)\right) - H_{1,\hat{\theta}_n}\left(H_{0,\hat{\theta}_n}^{-1}(q_0)\right)\right] \qquad (16) \\ &+ \sqrt{n}\left[H_{1,\hat{\theta}_n}\left(H_{0,\hat{\theta}_n}^{-1}(q_0)\right) - H_{1,\theta}\left(H_{0,\theta}^{-1}(q_0)\right)\right]. \qquad (17) \end{aligned}$$

Now we examine the asymptotic properties of (15) - (17) one at a time.

First we show that (15) $= o_P(1)$. Let θ be the true model parameters, so the MLE $\hat{\theta}_n \to_P \theta$, and $\sup_{t\in\mathbb{R}} |H_{1,\hat{\theta}_n}(t) - H_{1,\theta}(t)| \to_P 0$ due to the condition that the derivative of $H_{1,\theta}$ with respect to θ is uniformly bounded in θ. Suppose $g_1(x) = \mathrm{I}_{[x\le t_1]}$ and $g_2(x) = \mathrm{I}_{[x\le t_2]}$ $(\forall t_1, t_2 \in \mathbb{R})$ are two indicator functions from $\mathcal{F}$ and without loss of generality, $t_1 < t_2$, then under the $L_2(P)$-metric as the semimetric on $\mathcal{F}$, namely,

$$\begin{aligned}\rho_{P_n}(g_1, g_2) &= \sqrt{\int (g_1 - g_2)^2 dP_n} = \sqrt{\int \left(\mathrm{I}_{[t_1\le x\le t_2]}\right)^2 dH_{1,\hat{\theta}_n}(x)} \\ &= \sqrt{H_{1,\hat{\theta}_n}(t_2) - H_{1,\hat{\theta}_n}(t_1)},\end{aligned}$$

and

$$\begin{aligned}\rho_{P_0}(g_1, g_2) &= \sqrt{\int (g_1 - g_2)^2 dP_0} = \sqrt{\int \left(\mathrm{I}_{[t_1\le x\le t_2]}\right)^2 dH_{1,\theta}(x)} \\ &= \sqrt{H_{1,\theta}(t_2) - H_{1,\theta}(t_1)},\end{aligned}$$

we can show that,

$$\begin{aligned}\sup_{g_1,g_2\in\mathcal{F}} |\rho_{P_n}(g_1, g_2) - \rho_{P_0}(g_1, g_2)| &= \sup_{t_1,t_2\in\mathbb{R}} \left| \sqrt{H_{1,\hat{\theta}_n}(t_2) - H_{1,\hat{\theta}_n}(t_1)} - \sqrt{H_{1,\theta}(t_2) - H_{1,\theta}(t_1)} \right| \\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

This justifies Lemma 1.

Similarly, it is easily shown that $\sup_{t\in\mathbb{R}} |H_{0,\hat{\theta}_n}(t) - H_{0,\theta}(t)| \to_P 0$, by Lemma 3,

$$\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right) \to_d N\left(0, q_0(1-q_0)/h^2_{0,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right),$$

where $h_{0,\theta}$ is the density function of $H_{0,\theta}$. This implies that $\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0) = o_P(1)$, and hence

$$\int \left(\mathrm{I}_{[x\le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)]} - \mathrm{I}_{[x\le H^{-1}_{0,\theta}(q_0)]}\right)^2 dH_{1,\theta_0}(x) = \left| H_{1,\theta}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right) \right| \to_P 0,$$

by the continuous mapping theorem.

Let $f_n(x) = \mathrm{I}_{\left[x\le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right]}$, and $f_0(x) = \mathrm{I}_{\left[x\le H^{-1}_{0,\theta}(q_0)\right]}$. Therefore, by Lemma 2,

$$\begin{aligned}&\sqrt{m}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \\ &= \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^m \mathrm{I}_{\left[Y_{n,i}\le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right]} - \int \mathrm{I}_{[x\le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)]} dH_{1,\hat{\theta}_n}(x)\right] \\ &= \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^m f_n(Y_{n,i}) - \int f_n(x) dH_{1,\hat{\theta}_n}(x)\right] = \mathbb{G}_{n,P_n} f_n \rightsquigarrow \mathbb{G}_{P_0} f_0 = O_P(1).\end{aligned}$$

Accordingly,

$$\begin{aligned}(15) &= \sqrt{n/m}\sqrt{m}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \\ &= \sqrt{n/m}O_P(1) = o_P(1),\end{aligned}$$

since $\lim_{n\to\infty}(n/m) = 0$.

Next we show that (16) is also $o_P(1)$.

$$\begin{aligned}(16) &= \sqrt{\frac{n}{m}}\sqrt{m}\left[H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \\ &= \sqrt{\frac{n}{m}}\sqrt{m}\left[h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right. \\ &\qquad \left. + O_P\left(\left|\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right|^2\right)\right] \\ &= \sqrt{\frac{n}{m}}\left[h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right. \\ &\qquad \left. + \sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)^2 O_P(1)\right] \\ &= \sqrt{\frac{n}{m}}\left(O_P(1) + o_P(1)\right) = o_P(1),\end{aligned}$$

since

$$h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right) \to_d N\left(0, q_0(1-q_0)\frac{h^2_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)}{h^2_{0,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)}\right),$$

by Lemma 3 and

$$\begin{aligned}\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)^2 &= \left[\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right]\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right) \\ &= O_P(1)\cdot o_P(1) = o_P(1).\end{aligned}$$

Finally,

$$\begin{aligned}(17) &= \sqrt{n}\left[H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right] \\ &= \sqrt{n}\left\{H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right) - H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right) + H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right) - H_{1,\theta_0}\left(H^{-1}_{0,\theta}(q_0)\right)\right\} \\ &= \sqrt{n}\left\{h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right)\left(H^{-1}_{0,\hat{\theta}_n}(q_0) - H^{-1}_{0,\theta}(q_0)\right) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)(\hat{\theta}_n - \theta)\right\} + o_P(1) \\ &= \sqrt{n}\left\{h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right)\nabla_\theta H^{-1}_{0,\theta}(q_0)(\hat{\theta}_n - \theta) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)(\hat{\theta}_n - \theta)\right\} + o_P(1) \\ &= \left\{h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right)\nabla_\theta H^{-1}_{0,\theta}(q_0) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right\}\sqrt{n}(\hat{\theta}_n - \theta) + o_P(1).\end{aligned}$$

By the continuous mapping theorem and the delta method, (17) is asymptotically normal with mean 0 and covariance matrix $V = A\mathcal{I}^{-1}A^T$, where

$$A = h_{1,\theta}\left(H_{0,\theta}^{-1}(q_0)\right)\nabla_\theta H_{0,\theta}^{-1}(q_0) + \nabla_\theta H_{1,\theta}\left(H_{0,\theta}^{-1}(q_0)\right). \tag{18}$$

In summary, as $n \to \infty$,

$$\begin{aligned}
\sqrt{n}(\widehat{\text{sen}}_B - \text{sen}_B) &= \sqrt{n}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}_{0,\hat{\theta}_n,m}^{-1}(q_0)\right) - H_{1,\theta}\left(H_{0,\theta}^{-1}(q_0)\right)\right] \\
&= (15) + (16) + (17) \\
&= o_P(1) + o_P(1) + \sqrt{n}\left[H_{1,\hat{\theta}_n}\left(H_{0,\hat{\theta}_n}^{-1}(q_0)\right) - H_{1,\theta}\left(H_{0,\theta}^{-1}(q_0)\right)\right] \\
&\to_d N(0, V).
\end{aligned}$$

Appendix B Proof of Theorem 2

Since F_1 and F_0 are the CDF of bivariate normal distributions, the function $G(\mathbf{C}, \theta)$ is continuously differentiable with respect to $\mathbf{C}$ and θ. Condition (13) is equivalent to the statement that the matrix $\nabla_{\mathbf{C}} G(\mathbf{C}_0, \theta)$ is invertible, hence, according to the implicit function theorem (Kudryavtsev, 2001), there exists an open set U containing θ, an open set V containing $\mathbf{C}_0$, and a unique continuous differentiable function $g : U \to V$ such that $\mathbf{C} = g(\theta)$ and $G(g(\theta), \theta) = 0$ for all $\theta \in U$.

Based on the MLE properties, it is known that $\hat{\theta}_n \to_p \theta$ and $\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \mathcal{I}^{-1})$. So for any $\epsilon > 0$ and $\delta > 0$, there exists an N, such that $n > N$, $\Pr(|\hat{\theta}_n - \theta| > \delta) < \epsilon$. This implies that for any $n > N$, $\hat{\theta}_n \in U$ in probability, and hence the proposed method for finding the cut-off $\hat{\mathbf{C}}_n = (\hat{C}_{n,1}, \hat{C}_{n,2})$ through solving for $G(\hat{\mathbf{C}}_n, \hat{\theta}_n) = 0$ results in $\hat{\mathbf{C}}_n = g(\hat{\theta}_n)$ in probability.

Further note that $F_1(\mathbf{C}, \theta) = F_1(g(\theta), \theta)$ is a continuously differentiable function of θ, and consequently, by the continuous mapping theorem and the delta method, we have

$$\begin{aligned}
\sqrt{n}\,(\widehat{\text{sen}}_C - \text{sen}_C) &= \sqrt{n}\left(F_1(\hat{\mathbf{C}}_n, \hat{\theta}_n) - F_1(\mathbf{C}_0, \theta)\right) \\
&= \sqrt{n}\left(F_1(\hat{\mathbf{C}}_n, \hat{\theta}_n) - F_1(\mathbf{C}_0, \hat{\theta}_n) + F_1(\mathbf{C}_0, \hat{\theta}_n) - F_1(\mathbf{C}_0, \theta)\right) \\
&= \sqrt{n}\left(\nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \hat{\theta}_n)(\hat{\mathbf{C}}_n - \mathbf{C}_0) + \nabla_\theta F_1(\mathbf{C}_0, \theta)(\hat{\theta}_n - \theta)\right) + o_p(1) \\
&= \sqrt{n}\left(\nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \hat{\theta}_n)\nabla_\theta g(\theta)(\hat{\theta}_n - \theta) + \nabla_\theta F_1(\mathbf{C}_0, \theta)(\hat{\theta}_n - \theta)\right) + o_p(1) \\
&= \left(\nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \hat{\theta}_n)\nabla_\theta g(\theta) + \nabla_\theta F_1(\mathbf{C}_0, \theta)\right)\sqrt{n}(\hat{\theta}_n - \theta) \\
&\to_d N(0, B\mathcal{I}^{-1}B^T),
\end{aligned}$$

where

$$B = \nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \theta)\nabla_\theta g(\theta) + \nabla_\theta F_1(\mathbf{C}_0, \theta). \tag{19}$$

References

DeLong, E., D. DeLong, and C.-P. D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics 44*(3), 837–845.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–38.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Greenhouse, S. and N. Mantel (1950). Evaluation of diagnostic tests. *Biometrics 6*, 399–412.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics 31*, 201–224.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology 143*(1), 29–36.

Hanley, J. A. and B. J. McNeil (1983). A method of comparing the area under the receiver operating characteristic curves derived from the same cases. *Radiology 148*(3), 839–843.

Hui, S. L. and X. H. Zhou (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research 7*(4), 354–370.

Kraemer, H. C. (1992). *Evaluating medical tests: objective and quantitative guidelines*. Sage Publications.

Kudryavtsev, L. (2001). Implicit function. In M. Hazewinkel (Ed.), *Encyclopaedia of Mathematics*. http://eom.springer.de/i/i050310.htm: Springer.

Lefrère, J. J., F. Roudot-Thoraval, L. Morand-Joubert, J. C. Petit, J. Lerable, M. Thauvin, and M. Mariotti (1999). Carriage of GB virus C/hepatitis G virus RNA is associated with a slower immunologic, virologic, and clinical progression of human immunodeficiency virus disease in coinfected persons. *Journal of Infectious Diseases 179*, 783–789.

Linnet, K. (1987). Comparison of quantitative diagnostic tests: Type I error, power, and sample size. *Statistics in Medicine 6*, 147–158.

McClish, D. (1987). Comparing the area under more than two independent ROC curves. *Medical Decision Making 7*(3), 149–155.

McIntosh, M. W. and M. S. Pepe (2002). Combining several screening tests: optimality of the risk score. *Biometrics 58*, 657–664.

McLinden, J. H., T. M. Kaufman, J. Xiang, Q. Chang, D. Klinzman, A. M. Engel, G. Hess, U. Schmidt, M. Houghton, and J. T. Stapleton (2006). Characterization of an Immunodominant Antigenic Site on GB Virus C Glycoprotein E2 That Is Involved in Cell Binding. *Journal of Virology 80*(24), 12131–12140.

Nelsen, R. B. (1999). *An Introduction to Copulas*. New York: Springer.

Neyman, J. and E. Pearson (1933). On the problem of the most efficient tests of statisitcal hypothesis. *Philosophical Transactions of the Royal Society of London, Series A 231*, 289–337.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Su, J. Q. and J. S. Liu (1993). Linear combination of multiple diagnostic markers. *Journal of the American Statistical Association 88*, 1350–1355.

Tacke, M., K. Kiyosawa, K. Stark, V. Schlueter, B. Ofenloch-Haehnle, G. Hess, and A. M. Engel (1997). Detection of antibodies to a putative hepatitis G virus envelope protein. *Lancet 349*, 318–320.

Thompson, M. (2003). Assessing the diagnostic accuracy of a sequence of tests. *Biometrics 4*(3), 341–351.

Tillmann, H. L., H. Heiken, A. Knapik-Botor, S. Heringlake, J. Ockenga, J. C. Wilber, B. Goergen, J. Detmer, M. McMorrow, M. Stoll, R. E. Schmidt, and M. P. Manns (2001). Infection with GB Virus C and reduced mortality among HIV-Infected patients. *New England Journal of Medicine 345*(10), 715–724.

Van der Bij, A. K., N. Kloosterboer, M. Prins, B. Boeser-Nunnink, R. B. Geskus, J. M. A. Lange, R. A. Coutinho, and H. Schuitemaker (2005). GB Virus C Coinfection and HIV-1 Disease Progression: The Amsterdam Cohort Study. *Journal of Infectious Diseases 191*(5), 678–685.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer series in statistics. Springer.

Williams, C. F., D. Klinzman, T. E. Yamashita, J. Xiang, P. M. Polgreen, C. Rinaldo, C. Liu, J. Phair, J. B. Margolick, D. Zdunek, G. Hess, and J. T. Stapleton (2004). Persistent GB Virus C Infection and Survival in HIV-Infected Men. *New England Journal of Medicine 350*(10), 981–990.

Wu, Y. (2010). *The Partially Monotone Tensor Spline Estimation of Joint Distribution Function with Bivariate Current Status Data*. Ph. D. thesis, University of Iowa.

Xiang, J., S. Wünschmann, D. J. Diekema, D. Klinzman, K. D. Patrick, S. L. George, and J. Stapleton (2001). Effect of coinfection with GB virus C on survival among patients with HIV infection. *New England Journal of Medicine 345*, 707–714.

Zhang, W., K. Chaloner, H. Tillmann, C. Williams, and J. Stapleton (2006). Effect of early and late GB virus C viraemia on survival of HIV-infected individuals: a meta-analysis. *HIV Medicine 7*(3), 173–180.

Zhou, X. H., D. K. McClish, and N. A. Obuchowski (2002). *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience.