Biostatistics (2017), **0**, 0, *pp.* 1–32
doi:10.1093/biostatistics/manuscript˙submission

The ROC Curve for Cohort Designs

Haben Michael*, Lu Tian, and Musie Ghebremichael

Department of Statistics, Stanford University

haben.michael@stanford.edu

SUMMARY

The receiver operating characteristic (ROC) curve is a commonly used graphical summary of the discriminative capacity of a thresholded continuous scoring system for a binary outcome. Estimation and inference procedures for the ROC curve are well studied in the cross-sectional setting. However, there is a paucity of research when both biomarker measurements and disease status are observed longitudinally. In a motivating example, we are interested in characterizing the value of longitudinally measured CD4 counts for predicting the presence or absence of a transient spike in HIV viral load, also time-dependent. The existing method neither appropriately characterizes the diagnostic value of only observed CD4 counts nor efficiently uses status history in predicting the current spike status. We propose to jointly model the binary status as a Markov chain and the biomarkers levels, conditional on the binary status, as an autoregressive process, yielding a dynamic scoring procedure for predicting the occurrence of a spike. Based on the resulting prediction rule, we propose several straightforward extensions of the ROC curve to the longitudinal setting and describe procedures for statistical inference. Lastly, extensive simulations have been conducted to examine the small sample operational characteristics of the proposed methods.

Key words: HIV/AIDS; longitudinal binary outcomes; longitudinal biomarker; predictive value; receiver operator

*To whom correspondence should be addressed.

characteristic (ROC) curve.

1. INTRODUCTION

The receiver operating characteristic (ROC) curve is a graphical summary of the discriminative capacity of a thresholded continuous scoring system for a binary outcome. The curve consists of pairs of true positive and false positive rates as the threshold is varied (Swets and Pickett, 1982). Although many alternative measures have been proposed (Pencina *and others*, 2008; Steyerberg *and others*, 2010; Uno *and others*, 2007, 2013), the ROC curve remains the most commonly used in practice. In medical research, ROC curves are often used to characterize the quality of a continuous biomarker as a diagnostic for binary statuses such as diseased versus non-diseased (Pepe, 2003; Zhou *and others*, 2009). Well-studied in the cross-sectional setting, the ROC curve has been generalized to settings where the outcome of interest is time to event (Heagerty *and others*, 2000; Zheng and Heagerty, 2004; Heagerty and Zheng, 2005) and where the biomarker is longitudinally measured (Foulkes *and others*, 2010; Liu and Albert, 2014). However, less research has considered both longitudinal biomarker measurements and binary statuses.

A motivating example is data gathered from the Yale Prospective Longitudinal Pediatric HIV Cohort. The cohort comprises 97 children born to HIV-infected mothers in the New Haven, Connecticut, area since 1985. Various measurements were taken on the participants every 2–3 months over the 10-year period 1996–2006. Among these measurements, we focus on a continuous biomarker, CD4+ lymphocyte count, as a predictor of a binary outcome, "blip" status, the presence or absence of a transient spike in viral load (Paintsil *and others*, 2008).

Let X_{ij} and Y_{ij} denote the biomarker value and a binary status of patient i at visit j, respectively, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. The values X_{ij} may be direct biological measurements, as in the motivating example, or they may be derived or composite quantities. To assess the predictive value of the longitudinal biomarker X_{ij} for predicting Y_{ij}, Liu and Wu (2003) and Liu *and others* (2005) propose a simple mixed

effect regression model (Breslow and Clayton, 1993)

$$g\{\text{pr}(Y_{ij} = 1 \mid X_{i1}, \cdots, X_{in_i})\} = \alpha_i + \beta_i X_{ij}, \tag{1.1}$$

where $g(\cdot)$ is the logit function and α_i and β_i are, respectively, the subject-specific random intercept and slope. Similar models are described in Foulkes *and others* (2010) and Albert (2012). The random vector $(\alpha_i, \beta_i)'$ is assumed to follow a Gaussian distribution

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left\{\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \mathbf{\Gamma}_0\right\}.$$

The ROC curve summarizing the diagnostic value of X_{ij} is then constructed based on pairs

$$\left\{(\hat{\alpha}_i + \hat{\beta}_i X_{ij}, Y_{ij}), i = 1, \cdots, n; j = 1, \cdots, n_i\right\},$$

where $(\hat{\alpha}_i, \hat{\beta}_i)'$ are estimates of the subject-specific random effects obtained from the observed data, for example,

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} = E\left\{\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \middle| X_{i1}, \cdots, X_{in_i}, Y_{i1}, \cdots, Y_{in_i}, \hat{\alpha}_0, \hat{\beta}_0, \hat{\mathbf{\Gamma}}_0\right\},$$

and $\hat{\alpha}_0$, $\hat{\beta}_0$ and $\hat{\Gamma}_0$ are maximum likelihood estimators for the corresponding population parameters.

While this approach is simple and intuitive, we mention several limitations. First, the parametric assumptions may be too restrictive for some applications. For example, as discussed below, the Yale pediatric HIV data suggests greater dependency among biomarkers and disease statuses nearer in time. Neither is accounted for by model (1.1), which is symmetric in time. Second, the subject-specific random effect estimate $\hat{\alpha}_i$ and $\hat{\beta}_i$, as defined above, are not available at visit j as the biomarker levels $(X_{i(j+1)}, \cdots, X_{in_i})$ and responses $(Y_{i(j+1)}, \cdots, Y_{in_i})$ are not yet observed. Third, the approach uses the same data both to fit the model and, by using the fitted biomarkers to construct the ROC curve, to assess the quality of the fit. One expects the assessment to overestimate the true diagnostic quality of the biomarker. Efforts to set aside a group of patients for validation after estimation encounter the difficulty that subject effect estimates for the validation patients are unavailable (Foulkes *and others*, 2010). Lastly and more conceptually, the notion of ROC curve stands to be refined in the context of longitudinal measurements of multiple patients.

In contrast to the cross-sectional setting, several useful ROC curves suggest themselves. For example, the predictive performance of the biomarker for a given patient, as determined by that patient's history, can be quite different from the predictive performance for the entire patient population. In the next section, we propose a general framework to address these limitations.

2. METHOD

We first note two properties desirable in a framework for assessing diagnostic performance in the longitudinal, multiple subject design under consideration. First, to assess the predictive performance of a biomarker, the biomarker should depend only on data that is available when a prediction is to be made. We adopt the vantage of a practitioner who has previous biomarker and status data for a patient, is confronted with a current biomarker for the patient, and must now predict current status. As predictions may be made at different times, the accuracy of the prediction and the associated ROC will depend on time, with the corresponding prediction depending only on patient history available at that time. Second, two types of prediction performance should be differentiated: that for an individual patient and that for a patient population. For the former, we target the performance of $\mathcal{F}(\mathcal{H}_{ij})$ as a predictor of Y_{ij}, where $\mathcal{F}(\mathcal{H}_{ij})$ is a continuous score summarizing the predictive information contained in history up to visit j of a given patient i, and $\mathcal{H}_{ij} = \{X_{i1}, \ldots, X_{ij}, Y_{i1}, \ldots, Y_{i(j-1)}\}$. For the latter, we are interested in the predictive performance of $\mathcal{F}(\mathcal{H}_{ij})$ in the entire patient population at a time j, that is, marginalizing across patients.

In the following, we first generalize the simple mixed effect model (1.1) and discuss the two types of predictive performance under the proposed model. As discussed further below, easy extensions lead to more sophisticated models allowing for more flexible prediction rules. We assume that the longitudinal biomarker levels X_{ij} follow an autoregressive process conditional on disease status $Y_{ij} \in \{0, 1\}$, which are

generated by a Markov chain as in, e.g., Azzalini (1994). Specifically, we assume that for the i^{th} patient

$$\begin{aligned} \text{pr}\{Y_{ij} = b | Y_{i(j-1)} = a, \mathcal{H}_{i(j-1)}\} &= p_{abi}, a, b \in \{0, 1\}, j = 2, \cdots, \\ X_{ij} | \left\{ Y_{ij} = a, \mathcal{H}_{ij} \right\} &= \theta_i^{(a)} + \rho_0 X_{i(j-1)} + \epsilon_{ij}, j = 1, \cdots, \\ \text{pr}(Y_{i1} = a) &= p_a \quad \text{and} \quad X_{i0} = 0, \end{aligned} \tag{2.2}$$

where

$$\rho_0 \in (-1, 1)$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_0^2), j = 1, \cdots, n_i,$$

$$\xi_i = \begin{pmatrix} \theta_i^{(0)} \\ \theta_i^{(1)} \\ g(p_{01i}) \\ g(p_{11i}) \end{pmatrix} \sim N \left\{ \begin{pmatrix} \theta_0 \\ \theta_1 \\ \mu_{01} \\ \mu_{11} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_\theta & 0 \\ 0 & \boldsymbol{\Sigma}_p \end{pmatrix} \right\}, i = 1, \ldots, n,$$

independently and identically, and $\eta_0 = (p_0, \theta_0, \theta_1, \mu_{01}, \mu_{11}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_p)$ are hyperparameters. This set of parametric distributions for the random effects is chosen in part for convenience as they permit the model parameters to be estimated using many standard statistical software packages. Specifically, $\{\theta_0, \theta_1, \boldsymbol{\Sigma}_\theta, \rho_0, \sigma_0^2\}$ can be estimated by fitting the linear mixed effects model (Laird and Ware, 1982)

$$X_{ij} = \theta_i^{(0)} + (\theta_i^{(1)} - \theta_i^{(0)}) Y_{ij} + \rho_0 X_{i(j-1)} + \epsilon_{i(j-1)}, j = 2, \cdots, n_i,$$

and $\{\mu_{01}, \mu_{11}, \boldsymbol{\Sigma}_p\}$ can be estimated by fitting the generalized mixed effects model

$$g\{\text{pr}(Y_{ij} = 1 \mid Y_{i(j-1)})\} = \mu_{01i} + (\mu_{11i} - \mu_{01i}) Y_{i(j-1)}, j = 2, \cdots, n_i,$$

where $\mu_{abi} = g(p_{abi}), a = 0, 1$ and $b = 0, 1$. In addition, p_0 can be estimated by the observed proportion across patients at the initial visit. More importantly, under this model, we may link Y_{ij} with the observed history at visit j via a random effects model

$$\begin{aligned} g\left\{\text{pr}\left(Y_{ij} = 1 \mid \mathcal{H}_{ij}, \xi_i\right)\right\} &= g\left\{\text{pr}\left(Y_{ij} = 1 \mid X_{i(j-1)}, X_{ij}, Y_{i(j-1)}, \xi_i\right)\right\} \\ &= \alpha_{i0} + \alpha_{i1} Y_{i(j-1)} + \alpha_{i2} X_{i(j-1)} + \alpha_{i3} X_{ij}, \end{aligned} \tag{2.3}$$

where

$$\begin{aligned}
\alpha_{i0} &= \frac{1}{2\sigma_0^2}\left\{(\theta_i^{(0)})^2 - (\theta_i^{(1)})^2\right\} + \mu_{01i},\\
\alpha_{i1} &= \mu_{11i} - \mu_{01i},\\
\alpha_{i2} &= \frac{1}{\sigma_0^2}\rho_0\left(\theta_i^{(0)} - \theta_i^{(1)}\right),\\
\alpha_{i3} &= -\frac{1}{\sigma_0^2}\left(\theta_i^{(0)} - \theta_i^{(1)}\right).
\end{aligned} \tag{2.4}$$

Thus model (2.2) generalizes model (1.1), which treats X_{ij} as the only informative element of the subject's history conditional on the subject's effect.

2.1 *Individual patient ROC curve*

To evaluate the predictive performance of the biomarker or score M_{ij} (or its history) for patient i at time j, we would like to consider the contrast of two survival functions: $\tilde{S}_{ij}^{(1)}(m)$ vs. $\tilde{S}_{ij}^{(0)}(m)$, where

$$\tilde{S}_{ij}^{(a)}(m) = \text{pr}\left(M_{ij} > m \mid Y_{ij} = a, \xi_i, \rho_0, \sigma_0\right), a = 0, 1,$$

and

$$M_{ij} = \alpha_{i0} + \alpha_{i1}Y_{i(j-1)} + \alpha_{i2}X_{i(j-1)} + \alpha_{i3}X_{ij}.$$

M_{ij} uses the available history, since under model (2.2), Y_{ij} is conditionally independent of the remaining history given M_{ij}. We may then use the ROC curve $\tilde{R}_{ij}(u) = \tilde{S}_{ij}^{(1)}\left\{\left(\tilde{S}_{ij}^{(0)}\right)^{-1}(u)\right\}$ or derived statistics such as the area under the ROC curve $\int_{\mathbb{R}} \tilde{S}_{ij}^{(1)}(x)d\left\{1 - \tilde{S}_{ij}^{(0)}(x)\right\}$ to summarize this contrast.

As α_{ij}, $j = 0, 1, 2, 3$, depend on the unknown subject-specific random effect ξ_i, the score M_{ij} is unavailable in practice. An ROC curve based on M_{ij} can only serve as a theoretical benchmark. We therefore estimate the random effect ξ_i based on its conditional distribution $\xi_i \mid \bar{\mathcal{H}}_{i(j-1)}$, where $\bar{\mathcal{H}}_{i(j-1)} = \{(Y_{ik}, X_{ik}), k = 1, \cdots, (j-1)\}$, and use a plug-in estimator for M_{ij} . For example, we may estimate the random effects and

M_{ij} by the posterior mean

$$\hat{\xi}_i(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) = \begin{pmatrix} \hat{\theta}_i^{(0)}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) \\ \hat{\theta}_i^{(1)}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) \\ g\left\{\hat{p}_{01i}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0)\right\} \\ g\left\{\hat{p}_{11i}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0)\right\} \end{pmatrix} = E\left\{ \begin{pmatrix} \theta_i^{(0)} \\ \theta_i^{(1)} \\ \mu_{00i} \\ \mu_{11i} \end{pmatrix} \middle| \, \bar{\mathcal{H}}_{i(j-1)}, \eta_0 \right\},$$

and

$$\hat{M}_j(\mathcal{H}_{ij}, \eta_0) = \alpha_{0j}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) + \alpha_{1j}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) Y_{i(j-1)}$$
$$+\alpha_{2j}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) X_{i(j-1)} + \alpha_{3j}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0) X_{ij},$$

respectively (Robinson, 1991), where the functions $\alpha_{kj}(\bar{\mathcal{H}}_{i(j-1)}, \eta_0), k = 0, 1, 2, 3$, are obtained by replacing all the relevant subject-specific random effects in (2.4) with their estimated counterparts based on $\bar{\mathcal{H}}_{i(j-1)}$. For example,

$$\alpha_{3j}(\bar{\mathcal{H}}_{ij}, \eta_0) = \frac{1}{\sigma_0^2}\left\{E(\theta_i^{(1)} \mid \bar{\mathcal{H}}_{i(j-1)}, \eta_0) - E(\theta_i^{(0)} \mid \bar{\mathcal{H}}_{i(j-1)}, \eta_0)\right\}.$$

Here, the subscript j is used to emphasize that the prediction of the subject-specific random effect is made at the visit j using information up to visit $j-1$. An explicit expression for this choice of $\hat{\xi}_i(\bar{\mathcal{H}}_{i(j-1)}, \eta_0)$ can be found in Appendix A of the Supplementary Material. Using the estimated score $\hat{M}_j(\mathcal{H}_{ij}, \eta_0)$ (or $\hat{M}_j(\mathcal{H}_{ij}, \hat{\eta}_0)$ if η_0 is unknown) to predict the disease status at visit j, the predictive performance of patient i's biomarker X_{ij} at visit j can be summarized by the ROC curve

$$ROC_{ij}(u \mid \eta_0) = ROC(u \mid \xi_i, \eta_0, j) = S_{ij}^{(1)}\left\{\left(S_{ij}^{(0)}\right)^{-1}(u)\right\}$$

where

$$S_{ij}^{(a)}(m) = \text{pr}\left(\hat{M}_j(\mathcal{H}_{ij}, \eta_0) > m \mid Y_{ij} = a, \xi_i, \eta_0\right), a = 0, 1,$$

is the subject- and visit-specific survival function of the estimated score. $ROC_{ij}(u \mid \eta_0)$ depends on the joint distribution of the random history $\mathcal{H}_{ij}$ and the response Y_{ij} and thus also on the subject-specific random effect ξ_i. Since we do not have a convenient analytic expression for $ROC(u|\xi, \eta_0, j)$, we resort to a Monte-Carlo method. Specifically, for the i^{th} patient:

1. Simulate $\mathcal{H}^*_{ij} = \{Y^*_{i1}, \cdots, Y^*_{i(j-1)}, X^*_{i1}, \cdots, X^*_{ij}\}$ and Y^*_{ij} according to model (2.2) using the subject specific random effect ξ_i and the population parameter $(p_0, \sigma_0, \rho_0)'$
2. Compute $\hat{M}_j(\mathcal{H}^*_{ij}, \eta_0)$ according to (2.5).
3. Repeat steps 1–2 a large number of times and calculate the empirical ROC curve $\hat{ROC}(u|\xi_i, \eta_0, j)$ of the resulting pairs $\left\{\hat{M}_j(\mathcal{H}^*_{ij}, \eta_0), Y^*_{ij}\right\}$.

$\hat{ROC}(u|\xi_i, \eta_0, j)$ can serve as an approximation to the subject specific ROC curve of the i^{th} patient at the j^{th} visit provided that this patient's subject-specific parameters are known or can be estimated up to the desired accuracy. When this assumption is unmet, e.g., when the time j is small and few observations on the patient of interest are available, we instead propose two alternative summaries of the diagnostic performance of the biomarker at the individual level.

The first is the average individual specific ROC curve over the patient population,

$$ROC_1(u|j) = E\left\{ROC(u|\xi, \eta_0, j)\right\}, \tag{2.5}$$

where the expectation is taken with respect to the random effect ξ. In practice, we may use Monte-Carlo methods, simulating a large number B of random effects $\{\tilde{\xi}_1, \cdots, \tilde{\xi}_B\}$ from the distribution for the random effect and estimating $ROC_1(u|j)$ by

$$\hat{ROC}_1(u|\eta_0, j) = B^{-1}\sum_{b=1}^{B}\hat{ROC}(u|\tilde{\xi}_b, \eta_0, j).$$

The resulting $\hat{ROC}_1(u|\eta_0, j)$ is not the ROC curve for any individual patient but the expected patient-level ROC curve for a typical patient from the given population. As before, when η_0 is unknown, we may replace it by a consistent estimator $\hat{\eta}$ and let

$$\hat{ROC}_1(u|j) = \hat{ROC}_1(u|\hat{\eta}, j).$$

Since $\hat{ROC}_1(u|\eta_0, j)$ is a smooth function of η_0, $\hat{ROC}_1(u|j)$ is a consistent estimator for $ROC_1(u|j)$ and $\sqrt{n}\left\{\hat{ROC}_1(u|j) - ROC_1(u|j)\right\}$ converges weakly to a mean zero Gaussian process indexed by u when $\sqrt{n}(\hat{\eta} - \eta_0)$ converges weakly to mean zero Gaussian distribution.

The second option is the limit

$$ROC_{2i}(u) = \lim_{j\to\infty} ROC(u|\xi_i, \eta_0, j). \tag{2.6}$$

As $j \to \infty$, $\hat{\xi}_i(\bar{\mathcal{H}}_{ij})$, $\hat{M}_j(\mathcal{H}_{ij})$ and $\text{pr}(Y_{ij} = a)$ converge to ξ_i, M_{ij} and π_{ai}, respectively, where

$$\pi_{ai} = \frac{p_{(1-a)ai}}{p_{(1-a)ai} + p_{a(1-a)i}}, a = 0, 1$$

are subject-specific state probabilities of the stationary distribution of the 2-state Markov chain. Therefore, provided $\alpha_{i3} \neq 0$,

$$\begin{aligned}
S^{(a)}_{i\infty}(m) &= \lim_{j\to\infty} S^{(a)}_{ij}(m) \\
&= \lim_{j\to\infty} \text{pr}\left\{\hat{M}_j(\mathcal{H}_{ij}) > m \mid Y_{ij} = a, \xi_i, \rho_0, \sigma_0, \right\} \\
&= \lim_{j\to\infty} \sum_{b=0,1} \text{pr}\left(X_{ij} - \rho_0 X_{i(j-1)} > \frac{m - \alpha_{i0} - \alpha_{i1} b}{\alpha_{i3}} \;\middle|\; Y_{ij} = a, Y_{i(j-1)} = b\right) \frac{p_{bai}\pi_{bi}}{\pi_{ai}} \\
&= \sum_{b=0,1} \left\{1 - \Phi\left(\frac{m - \alpha_{i0} - \alpha_{i1} b - \theta^{(a)}_i \sigma_0}{\sigma_0 \alpha_{i3}}\right)\right\} \frac{p_{bai}\pi_{bi}}{\pi_{ai}},
\end{aligned} \tag{2.7}$$

where $\Phi(\cdot)$ is cumulative distribution function of the standard normal. Here, we used the fact that under model (2.2), $X_{ij} - \rho_0 X_{i(j-1)}$ given $\{Y_{ij} = a\}$ is normally distributed with mean $\theta^{(a)}_i$ and variance σ^2_0. When $\alpha_{i3} = \sigma_0^{-2}\left(\theta^{(1)}_i - \theta^{(0)}_i\right) = 0$, i.e., a patient's diseased and non-diseased biomarker means are the same, the posterior probability of positive event status (2.3) reduces to

$$pr(Y_{ij} = 1 \mid \mathcal{H}_{ij}, \xi_i) = \frac{p_{Y_{i(j-1)}1i}}{p_{Y_{i(j-1)}0i} + p_{Y_{i(j-1)}1i}},$$

posterior probability of a 2-state markov chain. Consequently, the ROC curve summarizes the performance of a 2-state markov chain in predicting the next state in this case. This performance serves as a limiting case when α_{i3} becomes small in magnitude, the biomarkers cease to provide useful discrimination, and the patient's prior status carries all the information about current status.

$ROC_{2i}(u) = S^{(1)}_{i\infty}\left\{(S^{(0)}_{i\infty})^{-1}(u)\right\}$ can be viewed as the ROC curve for subject i after adequate follow-up and therefore reflects the personalized diagnostic value of the biomarker for the given patient. It may or may not be similar to the population counterpart described in the next section. $ROC_{2i}(u)$ can be estimated

by $\hat{ROC}_{2i}(u)$, which is the same as $ROC_{2i}(u)$ with ξ_i and η_0 being replaced by $\hat{\xi}_i = \hat{\xi}_i(\bar{\mathcal{H}}_{in_i}, \hat{\eta})$ and $\hat{\eta}$, respectively. Assuming that $n_i/n = r_0 \in (0,1)$ and $n \to \infty$, $\hat{ROC}_{2i}(u)$ is consistent and $\sqrt{n_i + n}\left\{\hat{ROC}_{2i}(u) - ROC_{2i}(u)\right\}$ converges to a mean zero Gaussian process. Therefore, the key assumption for estimating $ROC_{2i}(u)$ in practice is that n_i be sufficiently large to allow acceptable estimation of the individual-specific random effect. The resulting estimated ROC curve can then be used to characterize the diagnostic value of the biomarker for an individual patient after sufficient follow-up.

Inference for $ROC_1(u|j)$ and $ROC_{2i}(u)$ can be carried out with the parametric bootstrap. One simulates fresh data using the estimated population parameter $\hat{\eta}$ from model (2.2) and obtains $ROC_1^*(u|j)$ and $ROC_{2i}^*(u)$, the estimators for the corresponding ROC curves, from the simulated data. The empirical distributions $ROC_1^*(u|j) - \hat{ROC}_1(u|j)$ and $ROC_{2i}^*(u) - \hat{ROC}_{2i}(u)$ based on a large number of simulations serve as approximations to the distributions of $\hat{ROC}_1(u|j) - ROC_1(u|j)$ and $\hat{ROC}_{2i}(u) - ROC_{2i}(u)$, respectively. Point-wise confidence intervals of $ROC_1(u|j)$ and $ROC_{2i}(u)$ can be constructed along these lines.

REMARK 2.1 One may be interested in the diagnostic value of X_{ij} at the j^{th} visit given the past history $\bar{\mathcal{H}}_{i(j-1)}$. In this case, the ROC curve can be constructed based on the conditional survival function

$$\begin{aligned}
&S_{ij}^{(a)}(m \mid \bar{\mathcal{H}}_{i(j-1)}) \\
=&\text{pr}\left(X_{ij} > m \mid Y_{ij} = a, \xi_i, \bar{\mathcal{H}}_{i(j-1)}, \eta_0\right) \\
=&\text{pr}\left(X_{ij} > m \mid Y_{ij} = a, X_{i(j-1)}, \xi_i, \rho_0, \sigma_0\right) \\
=&1 - \Phi\left(\frac{m - \rho_0 X_{i(j-1)} - \sigma_0 \theta_i^{(a)}}{\sigma_0}\right).
\end{aligned}$$

In contrast to ROC curves based on M_{ij} or its estimator, this ROC curve reflects the predictive value of X_{ij} only. It also depends on the random effect $\theta_i^{(a)}$ unknown at the visit j. One may also consider its expectation with respect to random effects or its limit when $j \to \infty$ as an estimable alternative.

REMARK 2.2 Both $ROC_1(u|j)$ and $ROC_{2i}(u)$ are parametric in nature in that their summarization of the diagnostic value of the longitudinally measured biomarker are valid only if the parametric model (2.2) is

correctly specified.

2.2 *ROC curve for the patient population*

The predictive performance of the biomarker across the entire population may be very different from that for individual patient. For example, the latter does not take into account biomarker variation between patients, or differences between patients in the prior probabilities of positive status events. Were the data not longitudinal, we might consider the empirical ROC curve of biomarker–status pairs (X_i, Y_i), $i = 1, \ldots n$. To take accumulated patient data into account, we instead consider the ROC curve of $\hat{M}_j(\mathcal{H}_{ij}, \eta_0), i = 1, \ldots, n$, the patients' biomarker scores (2.5) at a given time j. The scores synthesize all the predictive information in the past history under model (2.2).

Conditionally on the population parameter η_0, the patient scores are iid, and the empirical ROC curve is a valid metric for predictive value of the scores regardless of validity of the model being used to derive them. If model (2.2) is a good approximation to the true relationship between $\mathcal{H}_{ij}$ and Y_{ij}, one may anticipate good prediction accuracy of the resulting score. A severely misspecified model may give a prediction score with poor performance. In either case, the ROC curve and derived statistics such as the area under the ROC curve remain objective measures for the predictive value of the scoring system.

Formally, assuming that $\lim_{n\to\infty} \hat{\eta} = \tilde{\eta}_0$ in probability, the score $\hat{M}_j(\mathcal{H}_{ij}, \hat{\eta})$ converges to

$$\begin{aligned}\hat{M}_j(\mathcal{H}_{ij}, \tilde{\eta}_0) = \alpha_{0j}(\bar{\mathcal{H}}_{i(j-1)}, \tilde{\eta}_0) + \alpha_{1j}(\bar{\mathcal{H}}_{i(j-1)}, \tilde{\eta}_0) Y_{i(j-1)} \\ + \alpha_{2j}(\bar{\mathcal{H}}_{i(j-1)}, \tilde{\eta}_0) X_{i(j-1)} + \alpha_{3j}(\bar{\mathcal{H}}_{i(j-1)}, \tilde{\eta}_0) X_{ij},\end{aligned}$$

where $\tilde{\eta}_0 = \eta_0$ if the model is correctly specified. We are interested in estimating the ROC curve for the predictive value at the jth visit

$$ROC_3(u|j) = S_j^{(1)}\left\{\left(S_j^{(0)}\right)^{-1}(u)\right\}, \tag{2.8}$$

where

$$S_j^{(a)}(m) = \text{pr}\left(\hat{M}_j(\mathcal{H}_{ij}, \tilde{\eta}_0) > m \mid Y_{ij} = a\right), a = 0, 1.$$

We do so by plugging in the empirical survival function:

$$\hat{ROC}_3(u|j) = \hat{S}_j^{(1)}\left\{\left(\hat{S}_j^{(0)}\right)^{-1}(u)\right\} \tag{2.9}$$

where $N_{aj} = \sum_{i=1}^{n} I(Y_{ij} = a)$,

$$\hat{S}_j^{(a)}(m) = N_{aj}^{-1}\sum_{i=1}^{n} I\left\{\hat{M}_{ij}(\mathcal{H}_{ij}) > m\right\} I(Y_{ij} = a)$$

and $I(\cdot)$ is the event indicator function. Similarly, the area under the ROC curve, the concordance statistics, may be estimated as

$$\hat{A}_j = \frac{1}{N_{1j}N_{0j}}\sum_{i=1}^{N_{0j}}\sum_{k=1}^{N_{1j}} I\left\{\hat{M}_j(\mathcal{H}_{ij},\hat{\eta}) > \hat{M}_j(\mathcal{H}_{kj},\hat{\eta})\right\} I(Y_{ij} = 1)I(Y_{kj} = 0).$$

In Appendix B, we show that $\hat{ROC}_3(u|j)$ is a consistent estimator for $ROC_3(u|j)$ and the distribution of $\sqrt{n}\left\{\hat{ROC}_3(u|j) - ROC_3(u|j)\right\}$ converges to a mean zero Gaussian process under mild regularity conditions. The variance of $\hat{ROC}_3(u|j)$ can be approximated via an efficient resampling method. At each iteration, we first generate random weights $\{W_1, \cdots, W_n\}$ from the unit exponential distribution and estimate η_0 under model (2.2) with the i^{th} observation weighted by W_i. Denote the estimator by η^* and let

$$ROC_3^*(u|j) = \hat{S}_j^{(1)*}\left\{\left(\hat{S}_j^{(0)*}\right)^{-1}(u)\right\}$$

where

$$\hat{S}_j^{(a)*}(m) = \frac{\sum_{i=1}^{N} I\left\{\hat{M}_j(\mathcal{H}_{ij},\eta^*) > m\right\} I(Y_{ij} = a)W_i}{\sum_{i=1}^{N} I(Y_{ij} = a)W_i}.$$

Obtaining in this way a large number of realizations of $ROC_3^*(u|j)$, their empirical variance can be used to approximate that of $\hat{ROC}_3(u|j) - ROC_3(u|j)$. Similar resampling methods can be used to make inference on A_j, the area under the ROC curve at the j^{th} visit.

The predictive value of the biomarker in the entire population also varies with the visit j. With more visits and richer data observed, the predictive ability of the updated scoring system is expected to increase. We may study the trend of predictive value from visit 1 to J by simultaneously estimating $ROC_3(u|2), \cdots,$ and $ROC_3(u|J)$. It is not difficult to show that

$$\left\{\hat{ROC}_3(u|2) - ROC_3(u|2), \cdots, \hat{ROC}_3(u|J) - ROC_3(u|J)\right\}$$

can be approximated by a multivariate mean zero Gaussian distribution, based on which the joint inference for the predictive value at all visits of interest may be conducted simultaneously.

When the predictive value of the constructed scoring system only varies moderately from visit j_1 to j_2, i.e., $ROC_3(u|j)$, $j_1 \leqslant j \leqslant j_2$ are similar, it is tempting to estimate the ROC curve evaluating the average predictive value between these two visits. To this end, one may empirically construct a ROC curve as

$$\hat{ROC}_4(u|j_1, j_2) = \hat{S}^{(1)}_{j_1 j_2}\left\{\left(\hat{S}^{(0)}_{j_1 j_2}\right)^{-1}(u)\right\},$$

where

$$\hat{S}^{(a)}_{j_1 j_2}(m) = N_{a j_1 j_2} \sum_{i=1}^{n}\sum_{j=j_1}^{j_2} I\left\{\hat{M}_{ij}(\mathcal{H}_{ij}) > m\right\} I(Y_{ij} = a),$$

and $N_{a j_1 j_2} = \sum_{i=1}^{n}\sum_{j=j_1}^{j_2} I(Y_{ij} = a)$. Since it averages observations from multiple visits, $\hat{ROC}_4(u|j_1, j_2)$ can be substantially more stable than $\hat{ROC}_3(u|j)$. $\hat{ROC}_4(u|j_1, j_2)$ is a consistent estimator of

$$ROC_4(u|j_1, j_2) = S^{(1)}_{j_1 j_2}\left\{\left(S^{(0)}_{j_1 j_2}\right)^{-1}(u)\right\},$$

where

$$S^{(a)}_{j_1 j_2}(m) = \frac{\sum_{j=j_1}^{j_2} S^{(a)}_{j}(m)\mathrm{pr}(Y_{ij} = a)}{\sum_{j=j_1}^{j_2}\mathrm{pr}(Y_{ij} = a)},$$

a weighted average of $S^{(a)}_{j}(m)$. Statistical inference based on $\hat{ROC}_4(u|j_1, j_2)$ can be made by resampling methods similar to those previously described.

REMARK 2.3 Despite some similarities, $ROC_1(u|j)$ and $ROC_3(u|j)$ are quite different. The former is parametric and interpretable only when model (2.2) is correctly specified, while and latter is nonparametric in nature. The former, ignoring the differentiability in biomarkers across patients, tends to be smaller than the latter.

REMARK 2.4 The proposed ROC curves depend on the patient history $\mathcal{H}_j$ through the biomarkers estimates $\hat{M}_j(\mathcal{H}_j, \hat{\eta})$. We may consider other functions of $\mathcal{H}_j$ given by different statistical models of the response. More generally, one may consider a working regression model

$$\mathrm{pr}\left(Y_{ij} = 1|\mathcal{H}_j\right) = \mathcal{F}(Y_{i1}, \cdots, Y_{ij}, X_{i1}, \cdots, X_{i(j-1)}; \boldsymbol{\theta})$$

and construct the ROC curve based on

$$\left\{\left(\mathcal{F}(Y_{i1},\cdots,Y_{ij},X_{i1},\cdots,X_{i(j-1)};\hat{\boldsymbol{\theta}}),Y_{ij}\right),i=1,\cdots,n\right\},$$

where $\mathcal{F}(\cdot;\boldsymbol{\theta})$ is a parametric function of observed history and $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are model parameter and its appropriate estimator, respectively.

2.3 *Extension*

In model (2.2), we assume that (i) the underlying disease status follows a simple Markov chain, i.e., the distribution of Y_{ij} only depends on $Y_{i(j-1)}$; and (ii) the distribution of the biomarker level at visit j, X_{ij}, only depends on $X_{i(j-1)}$ and Y_{ij}; see Figure 1. There are several obvious extensions :

1. The distribution of X_{ij} depends on $(X_{i(j-1)},Y_{i(j-1)},Y_{ij})$ (Figure 2)

2. The distribution of Y_{ij} depends on $(Y_{i(j-1)},X_{i(j-1)})$ (Figure 3)

Adapting model (2.2) to the first setting, where the biomarker value depends not only on the current disease status but also the status at the previous visit, gives:

$$X_{ij}|(X_{i(j-1)},Y_{i(j-1)},Y_{ij})=\theta_i^{(Y_{i(j-1)}Y_{ij})}+\rho_0X_{i(j-1)}+\epsilon_{ij},\quad \epsilon_{ij}\sim\mathcal{N}(0,\sigma_0^2), \tag{2.10}$$

where $(\theta_{i1},\theta_{i2},\theta_{i3},\theta_{i4})'$ is the subject-specific random effect and ϵ_{ij} is independent $\mathcal{N}(0,\sigma_0^2)$. Under this model

$$\begin{aligned}g\left\{\text{pr}(Y_{ij}=1\mid\mathcal{H}_{ij},\xi_i)\right\}&=g\left\{\text{pr}(Y_{ij}=1\mid X_{i(j-1)},X_{ij},Y_{i(j-1)})\right\}\\&=\alpha_{i0}+\alpha_{i1}Y_{i(j-1)}+\alpha_{i2}X_{i(j-1)}+\alpha_{i3}X_{ij}+\alpha_{i4}X_{i(j-1)}Y_{i(j-1)}+\alpha_{i5}X_{ij}Y_{i(j-1)},\end{aligned}$$

where

$$\alpha_{i0} = \frac{(\theta_i^{(00)})^2 - (\theta_i^{(01)})^2}{2\sigma_0^2} + \mu_{01i}$$
$$\alpha_{i1} = -\frac{(\theta_i^{(11)})^2 - (\theta_i^{(01)})^2 - (\theta_i^{(10)})^2 + (\theta_i^{(00)})^2}{2\sigma_0^2} + (\mu_{11i} - \mu_{01i})$$
$$\alpha_{i2} = \frac{\rho_0(\theta_i^{(00)} - \theta_i^{(01)})}{\sigma_0^2}, \quad \alpha_{i3} = -\frac{(\theta_i^{(00)} - \theta_i^{(01)})}{\sigma_0^2}, \quad \alpha_{i4} = -\frac{\rho_0(\theta_i^{(11)} - \theta_i^{(01)} - \theta_i^{(10)} + \theta_i^{(00)})}{\sigma_0^2},$$
$$\alpha_{i5} = \frac{(\theta_i^{(11)} - \theta_i^{(01)} - \theta_i^{(10)} + \theta_i^{(00)})}{\sigma_0^2}.$$

Therefore, besides the terms in (2.3), model (2.10) leads to additional interaction terms $X_{i(j-1)}Y_{i(j-1)}$ and $X_{ij}Y_{i(j-1)}$ contributing to the prediction of the disease status at the jth visit, Y_{ij}.

For the second setting, we may assume that

$$P\{Y_{ij} = b | Y_{i(j-1)} = a, X_{i(j-1)}\} = p_{abi}(X_{i(j-1)}), a, b \in \{0, 1\},$$

where

$$\begin{pmatrix} g\{p_{01i}(X_{i(j-1)})\} \\ g\{p_{11i}(X_{i(j-1)})\} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_{01i} \\ \mu_{11i} \end{pmatrix} + \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} X_{i(j-1)}, \mathbf{\Sigma}_p \right\}.$$

In other words, the transition probability of the underlying disease status depends on the biomarker level at the prior visit. Under this model

$$g\left\{\text{pr}(Y_{ij} = 1 \mid \mathcal{H}_{ij}, \xi_i)\right\} = g\left\{\text{pr}(Y_{ij} = 1 \mid X_{i(j-1)}, X_{ij}, Y_{i(j-1)})\right\}$$
$$= \alpha_0 + \alpha_{i1}Y_{i(j-1)} + \alpha_{i2}X_{i(j-1)} + \alpha_{i3}X_{ij} + \alpha_{i4}X_{i(j-1)}Y_{i(j-1)},$$

where

$$\alpha_{i0} = \mu_{01i} + \frac{(\theta_i^{(0)})^2 - (\theta_i^{(1)})^2}{2\sigma_0^2}$$
$$\alpha_{i1} = \mu_{11i} - \mu_{01i}$$
$$\alpha_{i2} = \frac{\rho_0(\theta_i^{(0)} - \theta_i^{(1)})}{\sigma_0^2} + \gamma_0, \quad \alpha_{i3} = -\frac{\theta_i^{(0)} - \theta_i^{(1)}}{\sigma_0^2}, \quad \text{and} \quad \alpha_{i4} = \gamma_1 - \gamma_0.$$

Therefore, compared with (2.3), there is an additional interaction term $X_{i(j-1)}Y_{i(j-1)}$ contributing to the prediction of the disease status at the jth visit, Y_{ij}. There may be more extensive generalizations of model (2.3) such as the combination of extensions of (1) and (2) or higher order Markov chains for Y_{ij}. As mentioned in the previous sections, while the validity of individualized ROC curve depends on the correct model specification, the population-based ROC curve $ROC_3(u|j)$ and $ROC_4(u|j_1, j_2)$ can be constructed for scoring systems developed under different modeling assumptions and used to compare different models in terms of their predictive ability.

3. EXAMPLE

The goal of highly active antiretroviral therapy in the treatment of HIV is to keep a patient's CD4 count high and to suppress viral load. *CD4 count* measures immunosuppression, the risk of opportunistic infections and the strength of the immune system. *Viral load* is the amount of HIV in a sample, indicative of, among other things, transmission risk. Both are tested regularly in a typical treatment regimen, approximately every 3–6 months.

Even when therapy is effective and viral load is clinically categorized as suppressed, transient spikes in viral load, or "blips," are observed. The clinical significance of viral blips is not understood well. While some studies have reported that viral blips are of no clinical significance, others have reported an association between viral blips and virologic failure. The identification of the predictors of viral blips and the association between viral blips and CD4+ T-cell changes over time is the subject of ongoing research. (See Paintsil *and others* (2016) and references therein.)

We consider the accuracy of absolute CD4+ T-lymphocyte count as a predictor of blip status among children. We analyzed longitudinal data from HIV-infected children enrolled in the Yale Prospective Longitudinal Cohort study comprising 97 children born to HIV-infected mothers in the greater New Haven, Connecticut, area since 1985. The predictor CD4 count measures the number of CD4 cells/mm^3 of blood and the response blip status is defined as a viral load equal or exceeding 50 copies/ml. The average age

at enrollment is 6.7 years, the median number of visits/patient is 32.5, a median 1 visit/3 months. Figure 4a summarizes the dates of visits in the lifetimes of the subjects. Further details on the cohort and definitions used here can be found in Paintsil *and others* (2008) and the references therein. Eighty-one subjects remain after excluding those with fewer than 20 visits.

A crude indication of the value of CD4 as a predictor of blip status is given in Fig. 4b, taken from Paintsil *and others* (2016). Here, all CD4 observations belonging to the 18 patients with positive blip status 50% or more of the time is compared with the 86 patients with blip status in fewer than 50% of their observations. Despite the large overlap, there is a clear location shift between the two measures. Figure 5a plots the trajectories of CD4 color-coded by blip status for a representative sample of subjects. The long sequences of like colors even as CD4 fluctuates wildly suggests previous blip status as a predictor of future blip status, motivating the markov structure in model (2.2). Finally, Figure 5b is a heatmap of the empirical correlation matrix among CD4 measurements on the first 40 visits. The progressively decreasing correlation according as entries are farther from the diagonal accords with the weak dependency of the autocorrelative structure in model (2.2).

We apply the ROC estimation procedure described in Section 2 to the pediatric HIV data in order to assess the value of the past CD4 counts and blip statuses as a predictor of current blip status. The MLE $\hat{\rho}_0 = .49$ indicates strong autoregressive structure, as suggested by the heatmap. Furthermore, $(\hat{\theta}_0, \hat{\theta}_1)/\hat{\sigma}_0 = (2.29, 3.11)$ and $\{g^{-1}(\hat{\mu}_{00}), g^{-1}(\hat{\mu}_{11})\} = (.93, .73)$ confirm the predictive value of the CD4 count and immediately prior blip status, respectively. The resulting ROC curves and their associated 95% confidence intervals are summarized in Figure 6. Here the confidence intervals are constructed via bootstrap methods with 1000 bootstrap samples. Despite the noisy data presented in Figures 4b and 5a, the risk score taking into account both previous CD4 values and blip status performs reasonably well as a predictor of the current disease status. We also plot the time-asymptotic individual ROC curve ROC_{2i} for selected patients. The fourth patient exhibits a nonsmooth curve. The "elbow" arises when a patient's previous disease status is significantly more predictive than the patient's biomarkers of future disease status.

In such cases the ROC curve approximates the discrete behavior of a threshold predictor. The validity of aforementioned individualized ROC curves depends on the correct specification of model (2.2). If we view model (2.2) merely as a working device used to derive a risk score for predicting the blip status, we may use $ROC_3(u|j)$ as well as $ROC_4(u|j_1, j_2)$ to summarize the predictive value of the scoring system between visits j_1 and j_2. Due to the small sample size and infrequent occurrence of blip, we construct $ROC_4(u|6, 8)$ and its 95% confidence interval as shown in Figure 5a. The area under the ROC curve is 0.865 with a bootstrap standard error of .008, also indicating a good predictive value. The jagged shape of the ROC curve reflects the fact that few of the 81 patient scores lie in the overlap of the case and control distributions.

As a comparison, we also plotted the ROC curve by fitting the simple random effect model (1.1) based on scores $(\hat{\alpha}_i + \hat{\beta}_i X_{ij}, Y_{ij})$. As expected, the resulting ROC curves is substantially higher than $\hat{ROC}_3(u|j)$ or $\hat{ROC}_1(u|j)$. However, in using fitted values as scores it measures the quality of the theoretical fit of the data to the random effects model, rather than the quality of the biomarker as a predictor of the blip status.

4. SIMULATION

In this section, we investigate the finite sample performance of the proposed model. To this end, we first simulate observed data mimicking the HIV example. Specifically, $\{(X_{ij}, Y_{ij}), i = 1, \cdots, n, j = 1, \cdots, n_i\}$ are simulated via model (2.2) with the population parameter η_0 being the maximum likelihood estimator from the HIV data. First, we calculate $\hat{M}_j(\mathcal{H}_j, \eta_0)$ as a function of history $\mathcal{H}_j$ based on (2.5) and obtain an approximation to the true ROC curve $ROC_1(u|j) = E\{ROC(u|\xi, \eta_0, j)\}$ and $ROC_3(u|j)$ using the synthetic data with a sample size of 10^6. Second, for selected random effects ξ, we calculate $ROC_2(u|\xi) = \lim_{j\to\infty} ROC(u|\xi, \eta_0, j)$ based on the analytic expression of $S^{(a)}_{i\infty}(m)$ given in (2.7). The obtained ROC curves are presented in Figure 8.

We next generate data sets consisting of $n = 300$ patients with $n_i = 40$ visits each, again using as hyperparameters estimates obtained from the pediatric HIV data. For each simulated data set, we estimate

1. the expected individual-specific ROC curve $ROC_1(u|j)$ at $j = 3, 5, 9$ and its 95% point-wise confidence interval using the parametric bootstrap method;

2. the limiting individual-specific ROC curve $ROC_{2i}(u)$ for selected patients;

3. the population ROC curve $ROC_3(u|j)$ and its 95% point-wise confidence interval using the resampling method.

To evaluate the performance of the proposed method, we repeat the simulation 1000 times and estimate the empirical bias of the point estimators as well as the coverage level of the 95% confidence intervals at selected u for both $ROC_1(u|j)$ and $ROC_3(u|j)$, $j = 3, 5, 9$. The detailed simulation results for $u = 10\%, 25\%, 50\%$ and 75% are summarized at Table 1. The empirical biases are reasonably small in magnitude and the coverage level of the 95% confidence intervals are consistent with the nominal level allowed by the Monte-Carlo simulation error. In general, as expected, the population ROC curve tends to be higher than the individualized counterpart at the same visit. For estimating $ROC_{2i}(u)$, we compare the AUC under ROC curve, $\int_0^1 \hat{ROC}_2(u|\xi, \hat{\eta}, j)du$, based on data from increasing number of visits with the true limiting AUC value for selected random effects ξs. We are especially interested in whether the estimator converges to the truth as the number of visits increases under this correct model-specification. Figure 10 plots the number of visits against the difference between estimated AUCs and the truth with 5 different realizations of ξ, showing the expected convergence. However, the convergence is relatively slow and one needs to accumulate data from a large number of visits in order to achieve reasonable estimation accuracy.

In the second set of simulations, we examined the performance of the proposal under model misspecification. Specifically, we simulated the data from the random effect model (1.1). All the model-parameters are taken to be the maximum likelihood estimators from the HIV data. As we discussed above, the diagnostic value represented by the ROC curve based $\{(\hat{\alpha}_i + \hat{\beta}_i X_{ij}, Y_{ij}), i = 1, \cdots, n, j = 1, \cdots, n_i\}$ can't be achieved in practice but may serve as a benchmark. Since model (2.2) is misspecified, we focus on the population ROC curve only. First, we plot in Figure 9 the true $ROC_3(u|j)$, $j = 3, 5, 9$, and that based on

$(\hat{\alpha}_i + \hat{\beta}_i X_{ij}, Y_{ij})$ by setting $n = 10^6$. As expected, by comparison with the benchmark $ROC_3(u|j)$ fails to reflect the predictive value of the observed history due to model misspecification. We repeat the simulation with a sample size $n = 300$ and $n_i = 40$ to examine the empirical biases of the point estimators and empirical coverage levels of the 95% confidence intervals. Our expectation that the inference procedure for $ROC_3(u|j)$ remains valid in the presence of model mis-specification is borne out by the simulation results in Table (2).

5. DISCUSSION

We have proposed a set of ROC-based metrics and statistical methods for evaluating the predictive value of a biomarker in a longitudinal, multiple patient design. We emphasize three keys in extending the ROC curve from the cross-sectional to longitudinal setting: (i) the score used to construct the ROC curve should take into account all of the observed history; (ii) the score should not take into account unobserved history; and (iii) the predictive value of the biomarker at the individual and the population levels should be treated differently. These objectives are not met by the mixed effects model (1.1) available in the literature, where (i) a patient's observations are conditionally independent given the subject effects, and in particular past observations are not taken into account in using the observation as a score; (ii) all time points are used to estimate the subject effects, so that the score estimate for a given time point is a function of the disease status it is intended to predict; and (iii) there is no distinction between patient and population ROC curves.

The current approach is developed based on a simple parametric model. While the pertinent parametric assumptions are sensible, they are chosen mainly for convenience in implementation and motivated by the HIV data. Necessary model checking analysis for other data is needed.

In the proposed approach, we assume that the biomarker is measured at a regular time interval, which is true in the HIV example. However, in clinical practice, the measurement time is often irregular and it may not be possible to group the measurements into comparable 1st, 2nd ... visits. As discussed in the paper, the predictive value of the biomarker depends on the observed history, an important component of which

is the actual measurement times. Therefore, the irregular measurement schedule introduces complications in defining and studying the predictive value of the longitudinally measured biomarker. Further research in this direction is warranted.

6. Supplementary Material

Derivations of several formulas and facts used above may be found in the appendices of the Supplementary Material. Software in the form of R code and a sample input data set are available from the corresponding author or the online Supplementary Material.

Acknowledgments

The authors are grateful for financial support from *** . The authors also thank the study participants and the principal investigator of the study, *** , for sharing the data with us.

References

Albert, Paul S. (2012). A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statistics in Medicine* **31**(2), 143–154.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* **81**(4), 767–775.

Breslow, Norman E and Clayton, David G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**(421), 9–25.

Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*, Number v. 1 in Introduction to Time Series and Forecasting. Springer.

Foulkes, Andrea S, Azzoni, Livio, Li, Xiaohong, Johnson, Margaret A, Smith, Co-

LETTE, MOUNZER, KARAM AND MONTANER, LUIS J. (2010*a*). Prediction based classification for longitudinal biomarkers. *The annals of applied statistics* **4**(3), 1476.

FOULKES, A. S., AZZONI, L., LI, X., JOHNSON, M. A., SMITH, C., MOUNZER, K. AND MONTANER, L. J. (2010*b*, 2010 Sep). Prediction based classification for longitudinal biomarkers. *The annals of applied statistics* **4**, 1476–1497.

HEAGERTY, PATRICK J, LUMLEY, THOMAS AND PEPE, MARGARET S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**(2), 337–344.

HEAGERTY, PATRICK J AND ZHENG, YINGYE. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**(1), 92–105.

LAIRD, NAN M AND WARE, JAMES H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

LIU, DANPING AND ALBERT, PAUL S. (2014). Combination of longitudinal biomarkers in predicting binary events. *Biostatistics* **15**(4), 706–718.

LIU, HONGHU, LI, GANG, CUMBERLAND, WILLIAM AND WU, TONGTONG. (2005). Testing statistical significance of the area under a receiving operating characteristics ciurve for repeated measures design with bootstrapping. *Journal of Data Science* **3**, 257–278.

LIU, HONGHU AND WU, TONGTONG. (2003). Estimating the area under a receiver operating characteristic (roc) curve for repeated measures design. *Journal of Statistical Software* **8**, 1–18.

MCCULLAGH, P. AND NELDER, J.A. (1989). *Generalized Linear Models, Second Edition*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

PAINTSIL, ELIJAH, GHEBREMICHAEL, MUSIE, ROMANO, SOSTENA AND ANDIMAN, WARREN. (2008). Absolute CD4+ T-Lymphocyte count as a surrogate marker of pediatric HIV disease progression. *Pediatric Infectious Disease Journal* **7**, 629–635.

PAINTSIL, ELIJAH, MARTIN, RYAN, GOLDENTHAL, ARIEL, BHANDARI, SHREYA, ANDIMAN, WARREN AND GHEBREMICHAEL, MUSIE. (2016). Frequent episodes of detectable viremia in hiv treatment-experienced children is associated with a decline in CD4+ T-cells over time. *Journal of AIDS & Clinical Research* **7**, 565–577.

PENCINA, MICHAEL J, D'AGOSTINO, RALPH B AND VASAN, RAMACHANDRAN S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**(2), 157–172.

PEPE, MARGARET SULLIVAN. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, USA.

ROBINSON, GEORGE K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**(1), 15–32.

STEYERBERG, EWOUT W, VICKERS, ANDREW J, COOK, NANCY R, GERDS, THOMAS, GONEN, MITHAT, OBUCHOWSKI, NANCY, PENCINA, MICHAEL J AND KATTAN, MICHAEL W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21**(1), 128.

SWETS, J.A. AND PICKETT, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press Series in Cognition and Perception. Elsevier Science & Technology Books.

UNO, HAJIME, CAI, TIANXI, TIAN, LU AND WEI, LJ. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102**(478), 527–537.

UNO, HAJIME, TIAN, LU, CAI, TIANXI, KOHANE, ISAAC S AND WEI, LJ. (2013). A unified inference

procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine* **32**(14), 2430–2442.

ZHENG, YINGYE AND HEAGERTY, PATRICK J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**(4), 615–632.

ZHOU, XIAO-HUA, MCCLISH, DONNA K AND OBUCHOWSKI, NANCY A. (2009). *Statistical Methods in Diagnostic Medicine*, Volume 569. John Wiley & Sons.

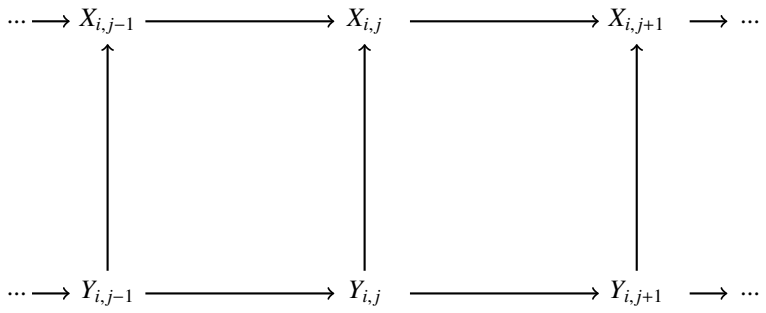


Figure 1: Model (2.2)

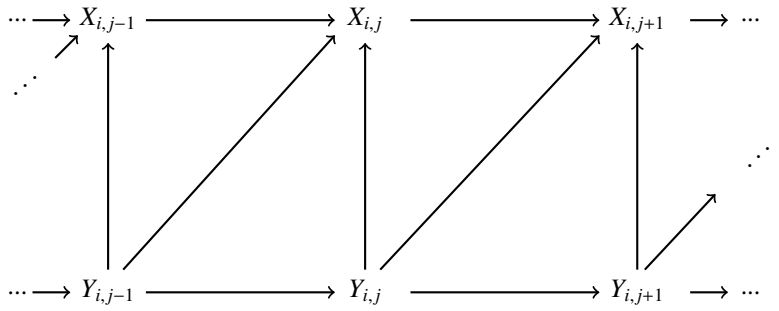


Figure 2: X_{ij} depends on $(X_{i(j-1)}, Y_{i(j-1)}, Y_{ij})$.

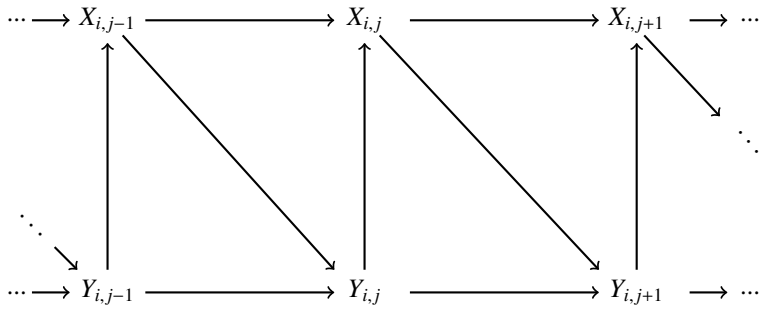


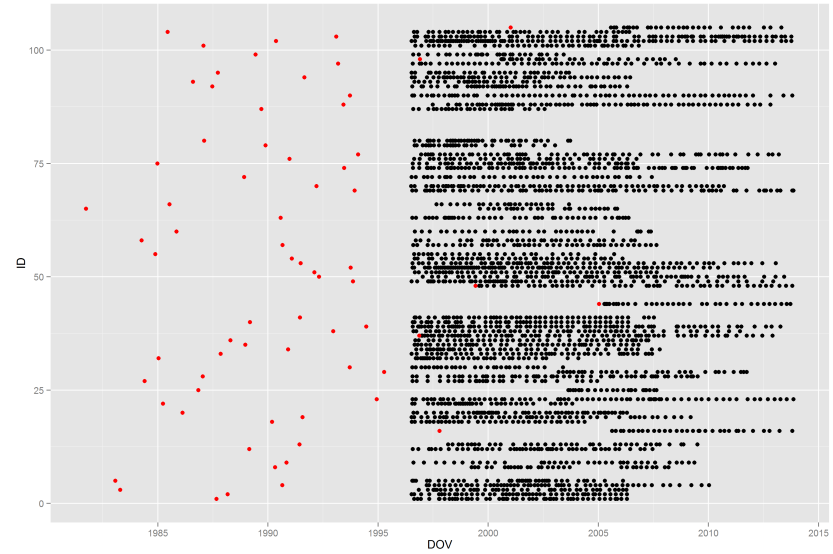
Figure 3: Y_{ij} depends on $(Y_{i(j-1)}, X_{i(j-1)})$.

Table 1: Nominal 95% CI coverage and bias of ROC_1 and ROC_3 for FPRs 10%, 25%, 50%, and 75% at visits 3, 5, and 7 (synthetic data using hyperparameters estimated from the pediatric HIV data, $N = 300$ patients).

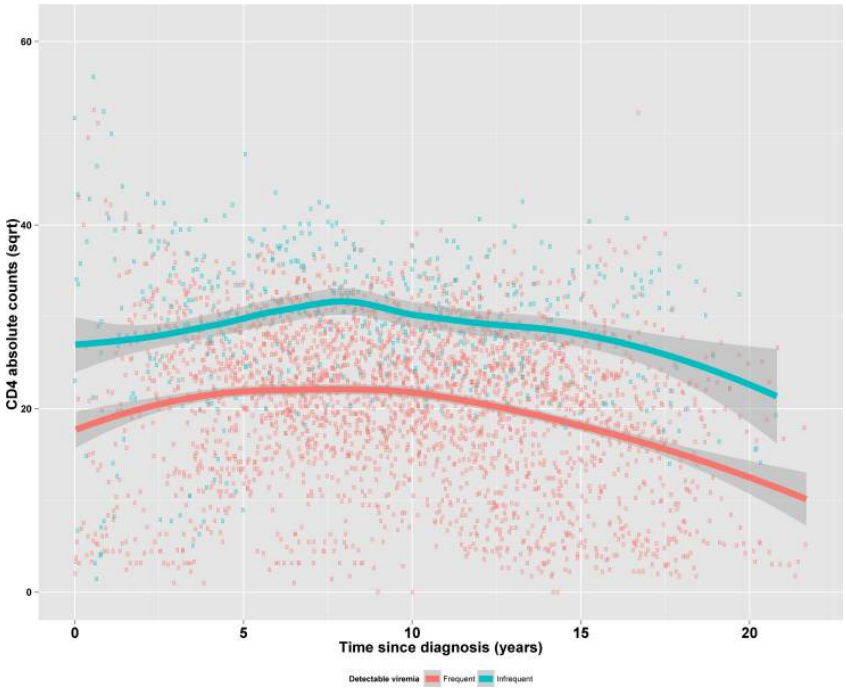
	FPR	0.10	0.25	0.50	0.75
t	**ROC**				
3	ROC_1	0.94 (0.034)	0.94 (0.038)	0.95 (0.028)	0.91 (0.022)
	ROC_3	0.94 (0.054)	0.93 (0.073)	0.92 (0.062)	0.94 (0.043)
5	ROC_1	0.96 (0.036)	0.96 (0.037)	0.92 (0.034)	0.92 (0.018)
	ROC_3	0.92 (0.064)	0.95 (0.064)	0.95 (0.052)	0.96 (0.035)
7	ROC_1	0.93 (0.039)	0.93 (0.052)	0.92 (0.029)	0.94 (0.017)
	ROC_3	0.95 (0.069)	0.94 (0.065)	0.94 (0.052)	0.95 (0.031)

Table 2: Misspecified model: Nominal 95% CI coverage and bias of $ROC_3(u|t)$ for FPRs 10%, 25%, 50%, and 75% at visits 3, 5, and 9 (random slope/intercept logistic model, $N = 300$ patients).

FPR	0.10	0.25	0.50	0.75
t				
3	0.95 (0.045)	0.91 (0.065)	0.94 (0.043)	0.96 (0.027)
5	0.94 (0.052)	0.93 (0.048)	0.92 (0.042)	0.93 (0.029)
9	0.94 (0.045)	0.97 (0.042)	0.93 (0.048)	0.94 (0.027)



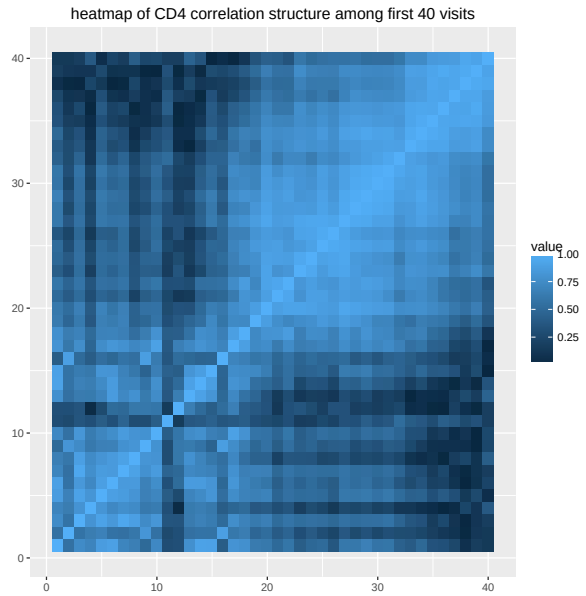
(a)



(b)

Figure 4: (a) Schedule of visits of patients enrolled in Yale pediatric HIV cohort. Horizontal series of dots correspond to a given patient's visits, black dots being visits and red dots indicating date of birth. The average age at enrollment is 6.7 years, the median number of visits/patient is 32.5, a median 1 visit/3 months. (b) Patients in the study were divided into two groups according as their proportion of positive blip observations exceeded 50% or not. The scatterplot of the observed CD4 values colored according to the group shows a clear location shift amid heavy overlap. (Figure taken from Paintsil *and others* (2016).)

(a)



(b)

Figure 5: (a) CD4 trajectory colored according to blip status. The (monochromatic) runs of a single status motivate the underlying markov structure in (2.2). (b) A heatmap of the empirical correlation of CD4 measurements for the first 40 visits suggests the autocorrelation structure in (2.2).

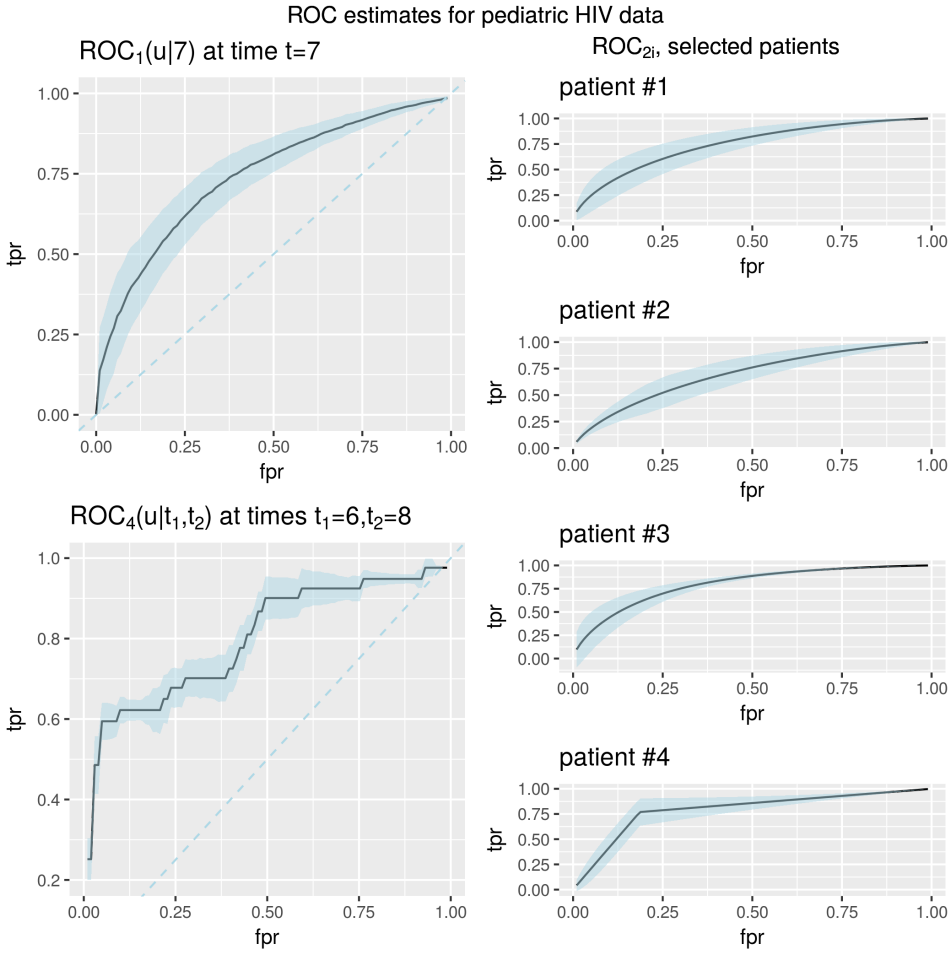


Figure 6: Expected individual ROC $ROC_1(u|t)$ and population ROC $ROC_4(u|t_1, t_2)$ at visits $t_1 = 6$ through $t_2 = 8$ with 95% bootstrap CI; limiting individual ROC ROC_{2i} for four patients (pediatric HIV data).

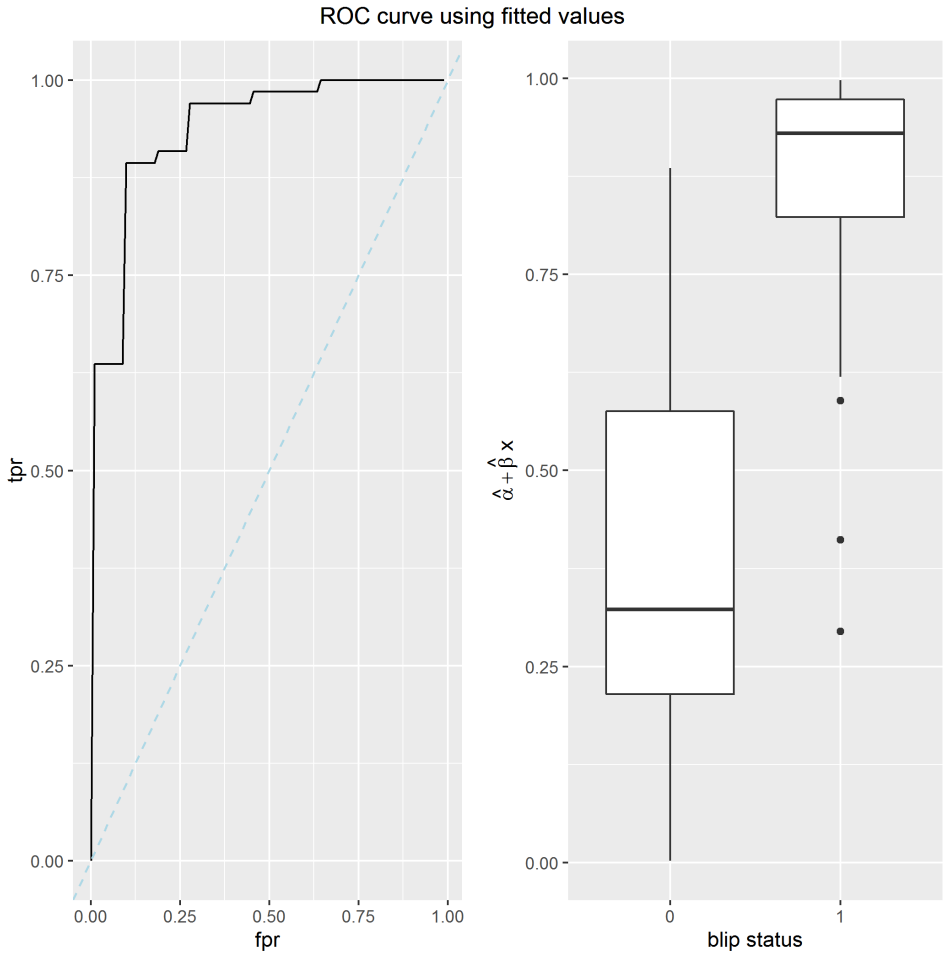


Figure 7: The ROC curve and boxplot when fitted values under a random effects model are used as scores, as in (1.1). The ROC curve may be overly optimistic as it does not account for generalization error of the predictor.

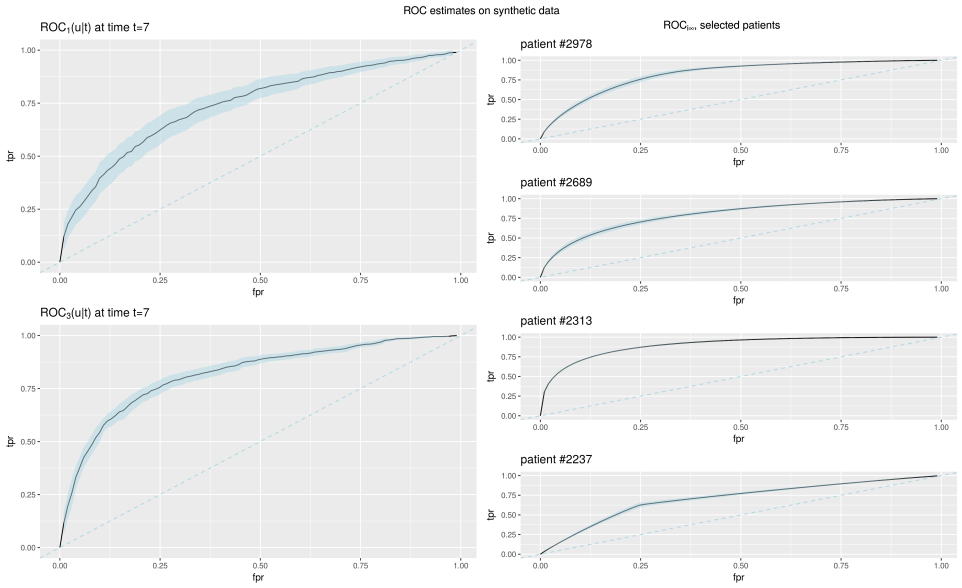


Figure 8: Expected individual ROC $ROC_1(u|t)$ and population ROC $ROC_3(u|t)$ at visit $t = 7$ with 95% bootstrap CI; limiting individual ROC $ROC_{i\infty}$ for four patients. The data was generated under model 2.2 with hyperparameters estimated from the pediatric HIV data.

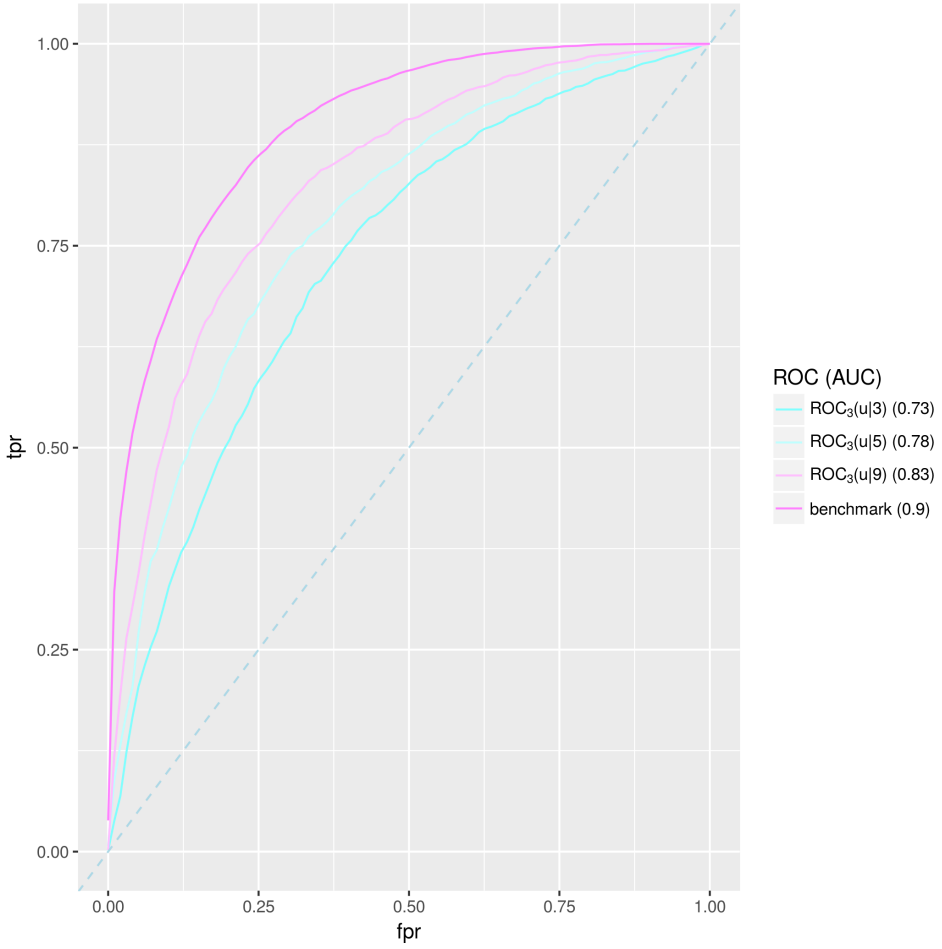


Figure 9: Misspecified model: Data is generated under the random effects model (1.1) and the population ROC, $ROC_3(u \mid t)$, is fit for visits $t = 3, 5, 9$. Also plotted is the true ROC curve for the model.

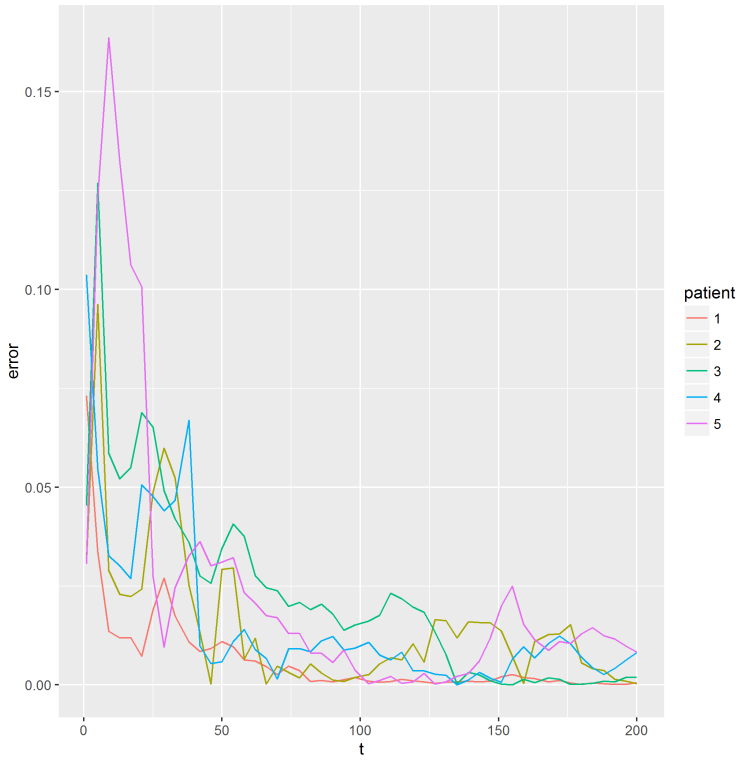


Figure 10: Convergence of estimates of limiting individual AUC for selected subject effects ξ as the number of visits t increases. The subject effects were sampled according to priors estimated from the pediatric HIV data.

Biostatistics (2017), **0**, 0, *pp.* 1–6
doi:10.1093/biostatistics/supplementary˙material

The ROC Curve in Cohort Design Supplementary Material

Lu Tian, Haben Michael*, [[Moses]]

Department of Statistics, Stanford University

haben.michael@stanford.edu

APPENDIX

A. Analytic expression of $\alpha_{ij}(\mathcal{H}_{ij}, \eta_0)$

Let $\vec{X}_{ij} = (X_{i1}, \ldots, X_{ij})$ and analogously for $\vec{\epsilon}_{ij}$. From the model

$$\begin{aligned} \text{pr}\{Y_{ij} = b | Y_{i(j-1)} = a, \mathcal{H}_{i(j-1)}\} &= p_{abi}, a, b \in \{0, 1\} \\ X_{ij} | \left\{ Y_{ij} = a, \mathcal{H}_{ij} \right\} &= \theta_i^{(a)} + \rho_0 X_{i(j-1)} + \epsilon_{ij} \\ \text{pr}(Y_{i1} = a) &= p_a \quad \text{and} \quad X_{i0} = 0, \end{aligned} \tag{A.1}$$

$\vec{X}_{ij}$ may be written

$$\vec{X}_{ij} = \begin{pmatrix} \theta_i^{(Y_{i1})} + \epsilon_{i1} \\ (\theta_i^{(Y_{i2})} + \epsilon_{i2}) + \rho_0(\theta_i^{(Y_{i1})} + \epsilon_{i1}) \\ \vdots \\ \sum_{k=1}^{j} \rho_0^{j-k}(\theta_i^{(Y_{ik})} + \epsilon_{ik}) \end{pmatrix} = \mathrm{P}_j \begin{pmatrix} \theta_i^{(Y_{i1})} + \epsilon_{i1} \\ \theta_i^{(Y_{i2})} + \epsilon_{i2} \\ \vdots \\ \theta_i^{(Y_{ik})} + \epsilon_{ik} \end{pmatrix} = \mathbf{P}_j \tilde{\mathbf{Y}}_{ij} \theta_i + \mathbf{P}_j \vec{\epsilon}_{ij},$$

where $\mathbf{P}_j$ is the $j \times j$ lower-left triangular matrix with entry m, n

$$(\mathbf{P}_j)_{mn} = \begin{cases} \rho_0^{m-n} & m \geqslant n \\ 0 & m < n \end{cases},$$

*To whom correspondence should be addressed.

$\tilde{\mathbf{Y}}_{ij}$ is the $j \times 2$ matrix consists of two j dimensional column vectors $1 - \vec{X}_{ij}$ and $\vec{X}_{ij}$ and $\theta_i = (\theta_i^{(0)}, \theta_i^{(1)})'$. Therefore,

$$\vec{X}_{ij} \sim \mathcal{N}(\mathbf{P}_j \tilde{\mathbf{Y}}_{ij} \theta_i, \mathbf{P}_j \mathbf{P}'_j).$$

Applying Bayes' formula, the posterior distribution of θ_i is

$$\mathcal{N}\left[(\tilde{\mathbf{Y}}'_{ij}\tilde{\mathbf{Y}}_{ij} + \mathbf{\Sigma}_\theta^{-1})^{-1}\left\{\tilde{\mathbf{Y}}'_{ij}\mathbf{P}_j^{-1}\vec{X}_{ij} + \mathbf{\Sigma}_\theta^{-1}(\theta_0, \theta_1)'\right\}, (\tilde{\mathbf{Y}}'_{ij}\tilde{\mathbf{Y}}_{ij} + \mathbf{\Sigma}_\theta^{-1})^{-1}\right].$$

Therefore

$$\left\{ \begin{array}{c} \theta_j^{(0)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) \\ \theta_j^{(1)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) \end{array} \right\} = (\tilde{\mathbf{Y}}'_{ij}\tilde{\mathbf{Y}}_{ij} + \mathbf{\Sigma}_\theta^{-1})^{-1}\left\{\tilde{\mathbf{Y}}'_{ij}\mathbf{P}_j^{-1}\vec{X}_{ij} + \mathbf{\Sigma}_\theta^{-1}(\theta_0, \theta_1)'\right\}.$$

Letting $n_{abi(j-1)} = \sum_{k=2}^{j-1} I(Y_{i(k-1)} = a)I(Y_{ik} = b), a = 0/1, b = 0/1$, Bayes rule again gives the posterior distribution for $\mu_i = (\mu_{01i}, \mu_{11i})' = \{g(p_{01i}), g(p_{11i})\}'$ with a density function of

$$\pi(\mu_i|\overline{\mathcal{H}}_{i(j-1)}) \propto \frac{\exp(\mu_{01i}n_{01ij} + \mu_{11i}n_{11ij})}{\{1 + \exp(\mu_{01i})\}^{n_{00i(j-1)}+n_{01i(j-1)}}\{1 + \exp(\mu_{11i})\}^{n_{10i(j-1)}+n_{11i(j-1)}}},$$

$$\times|2\pi\mathbf{\Sigma}_p|^{-1/2}\exp\left\{-\frac{1}{2}(\mu_{01i} - \mu_{01}, \mu_{11i} - \mu_{11})\,\mathbf{\Sigma}_p^{-1}\left(\begin{array}{c} \mu_{01i} - \mu_{01} \\ \mu_{11i} - \mu_{11} \end{array} \right)\right\}.$$

$\left\{\mu_{01j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0), \mu_{11j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0)\right\}$, the posterior mean of μ_i can be obtained accordingly and is a smooth function of $n_{abi(j-1)}$ and η_0. In summary,

$$\begin{aligned}
\alpha_{0j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) &= \frac{1}{2\sigma_0^2}\left[\{\theta_j^{(0)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0)\}^2 - \{\theta_j^{(1)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0)\}^2\right] + \mu_{01j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0), \\
\alpha_{1j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) &= \mu_{11j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) - \mu_{01j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0), \\
\alpha_{2j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) &= \frac{1}{\sigma_0^2}\rho_0\left\{\theta_j^{(0)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) - \theta_j^{(1)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0)\right\}, \\
\alpha_{3j}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) &= \frac{1}{\sigma_0^2}\left\{\theta_j^{(1)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0) - \theta_j^{(0)}(\overline{\mathcal{H}}_{i(j-1)}, \eta_0)\right\}.
\end{aligned}$$

B. Theoretical Properties of $\widehat{ROC}_3(u|j)$

In the derivation, we used the following Lemma:

Lemma B.1 Let $\mathcal{F} = \{f_\eta(x, s)\}_\eta$ be a class of measurable real-valued functions on $\mathcal{X} \times \mathcal{S}$ indexed by η, $\mathcal{S} = \{s_1, \ldots, s_K\}, K < \infty$. Let P denote the law of $(X, S) \in \mathcal{X} \times \mathcal{S}$, let $P(S = s_k) = p_k$ denote the

marginal probabilities of S, and let P_k denote the law of X conditional on $S = s_k$, extended to a law on $\mathcal{X} \times \mathcal{S}$ by setting $P_k(X, s_j) = 0$ for $j \neq k$. For a law Q on $\mathcal{X} \times \mathcal{S}$ let the semi-metric ρ_Q on $\mathcal{F}$ be given by $\rho_Q^2(f, g) = Var_Q(f(X, S) - g(X, S))$. If $\mathcal{F}$ is P_k-Donsker for $k = 1, \ldots, K$, and $(\mathcal{F}, \rho_P)$ is totally bounded, then $\mathcal{F}$ is P-Donsker.

Proof The P_k-Donsker property of $\mathcal{F}$ is equivalent (van der Vaart and Wellner (1996) 113–115) to the existence for any $\epsilon > 0$ of $n_k \in \mathbb{N}$, $\delta_k > 0$ such that for all $n > n_k$, $\delta < \delta_k$,

$$P(\sup_{\rho_{P|S=s_k}(f,g)<\delta} |\mathbb{G}(f) - \mathbb{G}(g)| > \epsilon \mid S = s_k) < \epsilon.$$

Let such $n_k, \delta_k, k = 1, \ldots K$, be given. The conditional variance formula $Var_P(f(X, S) - g(X, S)) = \sum_j p_j Var_{\mathbf{P}_j}(f(X, S) - g(X, S)) + Var_P(E_P(f(X, S) - g(X, S) \mid S))$ implies

$$\begin{aligned}
\min_j\{p_j\}\rho_{P_k}^2(f, g) &= \min_j\{p_j\} Var_{P_k}(f(X, s_k) - g(X, s_k)) \\
&\leqslant \sum_j p_j Var_{\mathbf{P}_j}(f(X, S) - g(X, S)) \\
&\leqslant Var_P(f(X, S) - g(X, S)) = \rho_P^2(f, g).
\end{aligned}$$

Then for $n > \max_j\{n_j\}$ and $\delta < \delta = \sqrt{\min_j\{p_j\}} \min_j\{\delta_j\}$,

$$\begin{aligned}
P(\sup_{\rho_P(f,g)<\delta} |\mathbb{G}(f) - \mathbb{G}(g)| > \epsilon) &= \sum_{k=1}^{K} p_k P(\sup_{\rho_P(f,g)<\delta} |\mathbb{G}(f) - \mathbb{G}(g)| > \epsilon \mid S = s_k) \\
&\leqslant \sum_{k=1}^{K} p_k P(\sup_{\rho_{P_k}(f,g)<\delta} |\mathbb{G}(f) - \mathbb{G}(g)| > \epsilon \mid S = s_k) < \epsilon,
\end{aligned}$$

which, given the total boundedness of $(\mathcal{F}, \rho_P)$, is equivalent to $\mathcal{F}$ being P-Donsker.

The first step is to show that the class of functions $\mathcal{F}_{aj} = \left\{ I\left\{\widehat{M}_j(\mathcal{H}, \eta) > m\right\} I(Y = a) \mid m, \eta \right\}$ is Donsker. From (A.1), we have

$$\begin{aligned}
\widehat{M}_j(\mathcal{H}, \eta) =& \frac{1}{2\sigma^2}\left[\{\theta_j^{(0)}(\mathcal{H}, \eta)\}^2 - \{\theta_j^{(1)}(\mathcal{H}, \eta)\}^2\right] + \mu_{01j}(\mathcal{H}, \eta) \\
&+ \left\{\mu_{11j}(\mathcal{H}, \eta) - \mu_{01j}(\mathcal{H}, \eta)\right\} Y_{i(j-1)} + \frac{1}{\sigma^2}\rho\left\{\theta_j^{(0)}(\mathcal{H}, \eta) - \theta_j^{(1)}(\mathcal{H}, \eta)\right\} X_{i(j-1)} \\
&+ \frac{1}{\sigma^2}\left(\theta_j^{(1)}(\mathcal{H}, \eta) - \theta_j^{(0)}(\mathcal{H}, \eta)\right) X_{ij}
\end{aligned}$$

and $\theta_j^{(a)}(\mathcal{H},\cdot)$ is linear in $\vec{\mathbf{x}}_j = (x_1,\cdots,x_j)'$. Therefore, for each of the 2^j possible values of $\vec{\mathbf{y}}_j = (y_1,\cdots,y_j)'$, $\widehat{M}_j(\mathcal{H},\eta)$ is a linear combination of $\{x_l, x_s x_t, 1 \leqslant l,s,t \leqslant j\}$. It follows Lemma 2.6.15 in van der Vaart and Wellner (1996) that the class of functions $\mathcal{F}_{\mathbf{y}_j} = \left\{ I\left\{\widehat{M}_j(h,\eta) > m\right\} \mid \nu,\eta \right\}$ is a VC class and in particular is Donsker for fixed $\vec{\mathbf{y}}_j$. Therefore, $\tilde{\mathcal{F}}_{\mathbf{y}_j} = \left\{ I\left\{\widehat{M}_j(h,\eta) > m\right\} I(y_j = a) \mid \nu,\eta \right\}$ is also Donsker. Lastly, by Lemma B.1, $\mathcal{F}_{aj}$ is P-Donsker (and thus also Glivenko Cantelli).

To show the consistency and asymptotical normality of the population ROC curve, we assume that $\widehat{\eta}$ converges in probability to $\tilde{\eta}_0 \in \Omega_\eta$, as $n \to \infty$ and

$$n^{-1/2}(\widehat{\eta} - \tilde{\eta}_0) = n^{-1/2} \sum_{i=1}^{n} \psi_i + o_p(1)$$

where $\tilde{\eta}_0$ is a deterministic vector, $E(\psi_i) = 0$ and $E(\psi_i^2) < \infty$. This condition is satisfied, when $\widehat{\eta}$ is obtained by solving a local linear estimating equation, whose limit has a unique solution Tian *and others* (2007). Here we don't need that the parametric model for (X_{ij}, Y_{ij}) is correctly specified. We also assume that Ω_M, the support of $\widehat{M}_j(\mathcal{H}_j,\eta)$, is bounded.

Let

$$\widehat{S}_{aj}(m,\eta) = n^{-1} \sum_{i=1}^{n} I\left\{\widehat{M}_{ij}(\mathcal{H}_{ij},\eta) > m\right\} I(Y_{ij} = a)$$

and

$$S_{aj}(m,\eta) = E\left\{ I\left\{\widehat{M}_j(\mathcal{H}_{ij},\eta) > v\right\} I(Y_{ij} = a)\right\}.$$

$$\begin{aligned}
&\sup_{m\in\Omega_M} |\widehat{S}_{aj}(m,\widehat{\eta}) - S_{aj}(m,\eta)| \\
\leqslant &\sup_{m\in\Omega_M} |\widehat{S}_{aj}(m,\widehat{\eta}) - S_{aj}(m,\widehat{\eta})| + \sup_{m\in\Omega_M} |S_{aj}(m,\widehat{\eta}) - S_{aj}(m,\tilde{\eta}_0)| \\
\leqslant &\sup_{(m,\eta)\in\Omega_M\times\Omega_\eta} |\widehat{S}_{aj}(m,\eta) - S_{aj}(m,\eta)| + \sup_{m\in\Omega_m} |S_{aj}(m,\widehat{\eta}) - S_{aj}(m,\tilde{\eta}_0)| \\
= &o_p(1),
\end{aligned}$$

where we used the Glivenko Cantelli property of $\mathcal{F}_{aj}$ and the continuity of $S_{aj}(m,\eta)$ in η at the last step. Therefore,

$$\widehat{S}_j^{(a)}(m) = \frac{\widehat{S}_{aj}(m,\widehat{\eta})}{n^{-1}\sum_{i=1}^{n} I(Y_{ij} = a)}$$

converges to

$$S_j^{(a)}(m) = \frac{S_{aj}(m, \tilde{\eta}_0)}{\text{pr}(Y_{ij} = a)}$$

uniformly in m. Next,

$$\begin{aligned}
& n^{1/2}\left\{\widehat{S}_{aj}(m,\widehat{\eta}) - S_{aj}(m,\eta)\right\} \\
=& n^{1/2}\left\{\widehat{S}_{aj}(m,\widehat{\eta}) - S_{aj}(m,\widehat{\eta})\right\} + n^{1/2}\left\{S_{aj}(m,\widehat{\eta}) - S_{aj}(m,\tilde{\eta}_0)\right\} \\
=& n^{1/2}\left\{\widehat{S}_{aj}(m,\tilde{\eta}_0) - S_{aj}(m,\tilde{\eta}_0)\right\} + n^{1/2}\left\{S_{aj}(m,\widehat{\eta}) - S_{aj}(m,\tilde{\eta}_0)\right\} + o_p(1) \\
=& n^{1/2}\left\{\widehat{S}_{aj}(m,\tilde{\eta}_0) - S_{aj}(m,\tilde{\eta}_0)\right\} + \dot{S}_{aj}(m,\tilde{\eta}_0) n^{-1/2} \sum_{i=1}^{n} \psi_i + o_p(1)
\end{aligned}$$

converges weakly to a mean zero Gaussian process indexed by $m \in \Omega_M$, where $\dot{S}_{aj}(m,\eta)$ is the partial derivative of $S_{aj}(m,\eta)$ with respect to η. Therefore,

$$\begin{aligned}
& n^{1/2}\left\{\widehat{S}_j^{(a)}(m) - S_j^{(a)}(m)\right\} \\
=& n^{1/2}\left[\frac{\widehat{S}_{aj}(m,\widehat{\eta})}{n^{-1}\sum_{i=1}^{n} I(Y_{ij} = a)} - \frac{S_{aj}(m,\tilde{\eta}_0)}{\text{pr}(Y_{ij} = a)}\right]
\end{aligned}$$

also converges to a mean zero Gaussian process index by $m \in \Omega_M$. Note that the functional $\mathcal{M}(S_1, S_2) = S_1\left\{(S_0)^{-1}(u)\right\}$ is differentiable in $\{S_1(\cdot), S_0(\cdot)\}$, by functional δ method,

$$\begin{aligned}
& n^{1/2}\left\{\widehat{R}_j(u) - \overline{R}_j(u)\right\} \\
=& n^{1/2}\left[\mathcal{M}\left\{\widehat{S}_j^{(1)}(\cdot), \widehat{S}_j^{(0)}(\cdot)\right\} - \mathcal{M}\left\{S_j^{(1)}(\cdot), S_j^{(0)}(\cdot)\right\}\right]
\end{aligned}$$

converges in distribution a mean zero Gaussian distribution Beutner and Zähle (2010). The justification of AUC under the ROC curve is similar by noting that the functional $\mathcal{A}\{S_1(\cdots), S_0(\cdot)\} = -\int S_1(u) dS_0(u)$ is also differentiable .

REFERENCES

BEUTNER, ERIC AND ZÄHLE, HENRYK. (2010). A modified functional delta method and its application to the estimation of risk functionals. *Journal of Multivariate Analysis* **101**(10), 2452–2463.

TIAN, LU, CAI, TIANXI, GOETGHEBEUR, ELS AND WEI, LJ. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**(2), 297–311.

VAN DER VAART, AAD W. AND WELLNER, JON A. (1996). *Weak Convergence and Empirical Processes*. Springer.