

Optimal Allocation of Gold Standard Testing under Constrained Availability: Application to Assessment of HIV Treatment Failure

Abstract

The World Health Organization (WHO) guidelines for monitoring the effectiveness of HIV treatment in resource-limited settings (RLS) are mostly based on clinical and immunological markers (e.g., CD4 cell counts). Recent research indicates that the guidelines are inadequate and can result in high error rates. Viral load (VL) is considered the “gold standard”, yet its widespread use is limited by cost and infrastructure. In this paper, we propose a diagnostic algorithm that uses information from routinely-collected clinical and immunological markers to guide a selective use of VL testing for diagnosing HIV treatment failure, under the assumption that VL testing is available only at a certain portion of patient visits. Our algorithm identifies the patient sub-population, such that the use of limited VL testing on them minimizes a pre-defined risk (e.g., misdiagnosis error rate). Diagnostic properties of our proposal algorithm are assessed by simulations. For illustration, data from the Miriam Hospital Immunology Clinic (RI, USA) are analyzed.

KEY WORDS: Antiretroviral failure, constrained optimization, HIV/AIDS, resource limited, ROC, tripartite classification.

Authors:

Tao Liu, Ph.D.
Joseph W. Hogan, Sc.D.
Lisa Wang, Sc.M.
Shangxuan Zhang, M.S.
Rami Kantor, M.D.

Author's Footnote:

Tao Liu is Assistant Professor (E-mail: tliu@stat.brown.edu), Joseph W. Hogan is Professor, and Lisa Wang is Graduate Student, Department of Biostatistics, Center for Statistical Sciences, Brown University School of Public Health, Providence, RI 02912. Shangxuan Zhang is Statistical Programmer, Memorial Sloan-Kettering Cancer Center, New York City, NY 10016. Rami Kantor is Associate Professor of Medicine, Division of Infectious Diseases, the Alpert Medical School of Brown University, Providence, RI 02912. This research is funded by a 2009 developmental grant from the Lifespan/Tufts/Brown Center for AIDS Research (CFAR). The project described is supported by Grant Number P30AI042853 from the National Institute of Allergy and Infectious Diseases (NIAID). Work by Dr. Kantor is also supported by a grant (Number R01AI66922) from the National Institute of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or NIH. The authors are grateful for the helpful comments from reviewers, the associate editor, and the editor. The authors also thank Ms. Allison K. DeLong for discussions and comments on early versions of the manuscript.

1. INTRODUCTION

According to a recent report of the World Health Organization (WHO) (WHO 2010a), almost 40 million people world-wide are infected with *human immunodeficiency virus* (HIV). Among them, over 97% live in resource-limited settings (RLS), particularly in sub-Saharan Africa (UNAIDS 2010). Although the number of people living with HIV remains high, the mortality rate due to acquired immune deficiency syndrome (AIDS) has started to decline since 2006 (UNAIDS 2009), due in large part to the successful rollout of HIV antiretroviral treatment (ART) in RLS (WHO 2010b).

With more and more people having access to ART, treatment failure is inevitable and must be anticipated. Treatment failure occurs when antiretroviral medications fail to control HIV replication in infected patients. Common causes of treatment failure include lack of proper medication adherence and development of drug resistance. The former may be addressed by reinforcing adherence (Gardner et al. 2009), while the latter usually mandates a switch to a more effective “next line” ART regimen (e.g., from a first- to a second-line regimen).

Monitoring the effectiveness of HIV treatment and correctly diagnosing treatment failure in a timely manner is critical for preventing HIV-related morbidity and mortality and transmission of the virus. Incorrect diagnosis of treatment failure can lead to undesired consequences and compromise the success that has been achieved by rolling out ART in RLS. Specifically, failure to diagnose treatment failure can result in continued viral replication, deterioration of patient’s immune system, extra clinical costs such as treatment of opportunistic infections, increased risk of HIV transmission, selection of resistant strains, and death (Anderson and Bartlett 2006; Calmy et al. 2007; Vekemans et al. 2007). Meanwhile, incorrectly diagnosing patients as having treatment failure when in fact they do not can prompt a premature switch to the next-line ART. This generates unnecessary financial burden (second-line therapies cost up to ten times more than first-lines) and potentially accelerates progression toward resistance to next-line therapies, which are most probably the last line in RLS (Vekemans et al. 2007).

In resource-rich countries such as those in much of western Europe and North America, viral load (VL) testing is routine for HIV treatment monitoring (Thompson et al. 2010; DHHS 2011). In this paper, VL refers to the amount of HIV in the blood as measured using nucleic acid amplification (Hammer et al. 2006). It is a marker that directly reflects the effectiveness of HIV treatment. Although HIV cannot be eradicated now, patients with adequate adherence can be expected to have *viral suppression*, which generally means that VL is below the lower detection limit of the assay being used (assays used for clinical purposes have lower detection limits of between 20 and 1000 copies/mL). A patient on adequate ART who has detectable VL after having previously reached an undetectable level is said to have *virological treatment failure* (hereafter “viral failure” or “treatment failure”), an indication that the particular treatment regimen may no longer be effective.

In RLS, VL testing is either limited or not available due to factors such as cost, lack of facilities, and lack of properly trained personnel (Fiscus et al. 2006; Calmy et al. 2007; Schooley 2007). Therefore, diagnosis of HIV treatment failure is commonly made using lower-cost and less accurate markers such as current CD4 cell count, CD4 percent among all lymphocytes, and relative changes in these measures since last visit; and clinical indicators

such as opportunistic infections, weight loss, and HIV-related malignancies. Indeed, these immunological and clinical markers form the basis of HIV treatment monitoring guidelines as recommended by the WHO (Calmy et al. 2007; WHO 2010a). These guidelines are widely adopted by countries in sub-Saharan Africa (e.g., Malawi 2003; Uganda 2003; Zambia 2004; Kenya 2005) and other developing regions.

Although CD4-based markers are generally associated with VL, a consensus has been reached recently that their use for diagnosing HIV treatment failure is prone to high misclassification rates (Deeks et al. 2000, 2002; Moore et al. 2005; Bisson et al. 2006; Schechter and Tuboi 2006; Tuboi et al. 2007; Bisson et al. 2008; Mee et al. 2008; Castelnuovo et al. 2009; Kantor et al. 2009; Keiser et al. 2009; Meya et al. 2009; Reynolds et al. 2009; Kiragga et al. 2012). Data from a recent study of patients receiving care through the Academic Model Providing Access to Healthcare (AMPATH) in western Kenya show that almost 40% of those having treatment failures would have been incorrectly diagnosed based on the WHO guidelines (Kantor et al. 2009).

Several studies have investigated monitoring HIV treatment using markers in addition to or instead of CD4 cell count (Bagchi et al. 2007; Kantor et al. 2009; Foulkes et al. 2010; Abouyannis et al. 2011). Bagchi et al. (2007) showed that weight loss is associated with treatment failure but pointed out that its clinical utility is limited because weight is influenced by many factors. Kantor et al. (2009) found in a Kenyan cohort that time on therapy and change in CD4 percent can be potentially incorporated into CD4-based rules to improve the diagnosis of treatment failure. Abouyannis et al. (2011) developed and tested a scoring system that incorporates CD4 count, mean cell volume, medication adherence, and HIV-associated clinical events for diagnosing treatment failure. Foulkes et al. (2010) proposed a prediction-based classification method that combines multiple time-varying clinical measures for predicting treatment failure. Each of these studies focuses on augmenting or replacing CD4 count with other immunological and clinical markers, assuming that VL testing is completely unavailable. Potential improvements are demonstrated, but often found to be marginal.

In this paper, we consider augmenting rules of diagnosing treatment failure based on low-cost markers (such as CD4 cell count) with a *selective use* of VL testing, under the assumption that VL testing can be ordered only for a fixed portion of patient visits. Our approach is motivated by the fact that several HIV care programs in developing countries have started to conduct VL testing for some of their patients. For example, as a result of the study by Kantor et al. (2009), AMPATH is currently conducting VL testing at about ten percent of its patient visits when treatment failure is suspected. Our approach is also motivated by the expectation that as technology and training advance (e.g., Greengrass et al. 2009), VL testing will be more affordable, even if substantially limited in the near future.

Assuming that VL testing is available but at a fixed portion of patient visits, we propose a tripartite classification procedure to triage VL testing based on a risk score S derived from low-cost non-VL markers. Specifically, the resulting tripartite diagnostic rule comprises two cut-off values l and u on S , with $l \leq u$, that classify HIV patients into three mutually exclusive categories (refer to Figure 1), and correspondingly takes one of the following three actions for each category.

- (a): Those with $S > u$ are diagnosed as failing treatment,

- (b): Those with $S \leq l$ are diagnosed as non-failing, and
- (c): Those with $l < S \leq u$ are designated for VL testing, which will provide an error-free diagnosis.

The tripartite diagnostic rule is designed to minimize a pre-specified risk (e.g., misclassification) subject to the constraint on the availability of VL assays. To identify the optimal rule, we develop both nonparametric and semiparametric approaches to inference about l and u . We also develop a receiver operating characteristic (ROC) analysis procedure for a general assessment of candidate tripartite rules. The ROC curve and the area under the ROC curve (AUC) provide a comprehensive measure of diagnosis capacity of tripartite rules, and allow us to evaluate the potential improvement that can be achieved by increasing VL testing availability. ROC analysis of tripartite rules has many statistical properties that are similar to conventional ROC analysis of bipartite rules.

The rest of the paper is organized as follows: Notations, definitions, and criteria for rule development are given in Section 2; nonparametric and semiparametric approaches to optimal rule selection are presented in Section 3; ROC analysis of tripartite rules is described in Section 4; and simulation studies are conducted in Section 5. For illustration, data from the HIV Immunology Clinic of the Miriam Hospital (RI, USA) are analyzed in Section 6. We conclude with a summary and discussion of future research in Section 7.

2. NOTATIONS AND DEFINITIONS

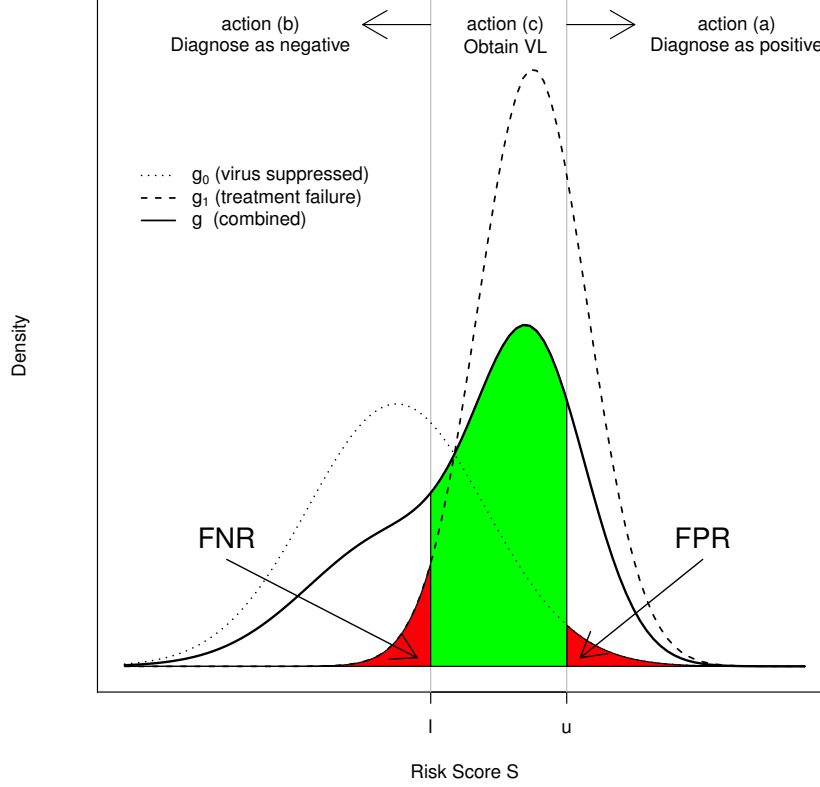
2.1 HIV Viral Status and Risk Score

The objective of HIV treatment monitoring is to diagnose viral failure. Let V denote a patient's (possibly unmeasured) viral load. Viral failure is said to occur when V exceeds a pre-specified threshold v^* , where v^* is typically the lower detection limit of the VL assay being used. Let $Z = \mathbf{1}(V > v^*)$ denote viral status with $Z = 1$ indicating a viral failure and 0 otherwise, where $\mathbf{1}(\cdot)$ is the indicator function. The prevalence of viral failure is denoted by $p = \Pr(Z = 1)$. At each patient encounter, a set of immunological, clinical, and demographic markers is usually collected, which may include CD4 count, CD4 percent among all lymphocytes (and recent changes in both), WHO stage, time on therapy, hemoglobin, weight, age, gender, and adherence measures. Henceforth, these markers are generically referred to as low-cost *clinical markers* and denoted by a vector \mathbf{X} .

For each individual, these clinical markers are translated into a scalar risk score $S = S(\mathbf{X})$. Several recent studies have proposed versions of $S(\mathbf{X})$ for determining the risk of treatment failure (e.g., Lynen et al. 2009; Meya et al. 2009; Abouyannis et al. 2011). If $S(\mathbf{X})$ is a predicted probability of viral failure given \mathbf{X} , it can be derived using logistic regression, regression trees, or other types of prediction-based classification methods (e.g., Pepe and Thompson 2000; Hastie et al. 2001; Foulkes et al. 2010; Justice et al. 2010; van der Laan 2011). In this paper, we assume that the functional form of $S(\cdot)$ is known, but note that finding and validating an optimal form of $S(\mathbf{X})$ is an important topic of research (see Huang et al. 2007; Pepe et al. 2008; Steyerberg et al. 2010; Pepe 2011).

Let G_1 and G_0 denote the distributions of S for patients with viral failure ($Z = 1$) and viral suppression ($Z = 0$), and g_1 and g_0 denote their associated densities, respectively. The population distribution of S is therefore a mixture distribution $G = (1 - p)G_0 + pG_1$, whose

Figure 1: Risk score distributions and diagnosis actions.



density is denoted by g . We assume that for independent observations S and S' , where $S \sim G_1$ and $S' \sim G_0$, S is stochastically greater than S' in the sense that on average, patients with viral failure have higher risk scores. An illustration of g_0 , g_1 , and g leading to a hypothetical distribution of S is presented in Figure 1.

2.2 Classification Cut-offs and Tripartite Rules

The tripartite diagnostic rule can be formalized as follows. Let l and u , with $l \leq u$, subdivide the population into three categories: those whose risk of treatment failure is high ($S > u$), low ($S \leq l$), or intermediate ($S \in \mathcal{J} \equiv (l, u]$). Let $\delta_{\mathcal{J}}(S)$ denote the diagnostic decision based on S , with $\delta_{\mathcal{J}}(S) = 1$ indicating a treatment failure diagnosis and $\delta_{\mathcal{J}}(S) = 0$ a non-failing diagnosis. Then our tripartite rule is expressed as

$$\delta_{\mathcal{J}}(S) = \begin{cases} 0 & \text{if } S \leq l, \\ Z & \text{if } S \in \mathcal{J}, \\ 1 & \text{if } S > u. \end{cases} \quad (1)$$

This rule obtains the gold standard measurement for the intermediate risk subpopulation $\{S \in \mathcal{J}\}$, which carries the greatest uncertainty about true viral status. Note that when

$S \in \mathcal{S}$, the diagnosis decision corresponds to the true viral failure status and therefore leads to a correct diagnosis.

2.3 Loss and Risk Functions

Let $L(d, z)$ denote the loss or cost incurred when the true viral failure status is $Z = z$ and a diagnostic decision d is taken. Two commonly used loss functions in studies of medical diagnosis are $L_1(d, z) = \mathbf{1}(d \neq z)$, which indicates whether a misdiagnosis occurs, and $L_2(d, z) = \{\mathbf{1}(d = 0, z = 1), \mathbf{1}(d = 1, z = 0)\}^\top$, which indexes misdiagnoses separately for those with viral failure (i.e., false negative, FN) and those without (i.e., false positive, FP). Loss functions can be made more elaborate and extended to incorporate potential costs as well as benefits of correct and incorrect diagnoses (e.g., expected mortality, cost of switching to next-line therapies, and gain of life expectancy); see Parmigiani (2002) for further discussions.

The development of our diagnostic rule also uses a weighted loss function

$$L_3(d, z; \lambda) = \lambda \mathbf{1}(d = 0, z = 1) + (1 - \lambda) \mathbf{1}(d = 1, z = 0),$$

where $\lambda \in [0, 1]$ is a user-specified weight that reflects relative loss for the two types of misdiagnoses. At the extremes, setting $\lambda = 1$ places the highest priority on avoiding FN (incorrectly diagnosing a patient as non-failing), while $\lambda = 0$ prioritizes avoidance of FP (incorrectly diagnosing a patient as treatment failure). An appropriate and meaningful value of λ should be contextually specific and take into account the available information about patient's health status and various costs associated with FP and FN.

The overall diagnostic accuracy of a diagnostic rule is summarized by a risk function defined as $R(\mathcal{S}) = \mathbb{E}[L(\delta_{\mathcal{S}}(S), Z)]$, where the expectation is taken over the joint distribution of $(S, Z)^\top$ (Berger 1985). For the loss function L_1 , $R_1(\mathcal{S}) = \mathbb{E}[\mathbf{1}\{\delta_{\mathcal{S}}(S) \neq Z\}]$ is the total misclassification rate (TMR). For L_2 , $R_2(\mathcal{S}) = \{p\text{FNR}, (1-p)\text{FPR}\}^\top$, where FNR and FPR are the FN and FP rates, respectively. For L_3 , we have a weighted sum of FPR and FNR

$$R_3(\mathcal{S}; \lambda) = \lambda p\text{FNR} + (1 - \lambda)(1 - p)\text{FPR}, \quad (2)$$

where the weights depend on both λ and the prevalence of viral failure. Risk function $R_3(\mathcal{S}; \lambda)$ is one form of 'net benefit' functions that have been used in decision curve analyses and utility analyses (Vickers and Elkin 2006; Baker 2009). As a special case when $\lambda = .5$, minimizing $R_3(\mathcal{S}; .5)$ is equivalent to minimizing $R_1(\mathcal{S})$.

In the next section, we develop methods for obtaining optimal rules under the risk criteria $R_1(\mathcal{S})$ and $R_3(\mathcal{S}; \lambda)$. The optimal rules that minimize $R_1(\mathcal{S})$ and $R_3(\mathcal{S}; \lambda)$ are called the min-TMR rules and min- λ rules, respectively. In Section 4, the vector-valued risk function $R_2(\mathcal{S})$ is used to develop a ROC analysis procedure for a general assessment of tripartite diagnostic rules.

3. OPTIMAL RULE SELECTION: CONSTRAINED OPTIMIZATION

3.1 Characterization of Constraints on Gold Standard Testing

Suppose that VL tests can be ordered for a fixed portion ϕ of patient visits, where $0 \leq \phi \leq 1$. Then the proposed tripartite rules must satisfy the constraint

$$G(u) - G(l) \leq \phi. \quad (3)$$

In the extreme cases, $\phi = 0$ means that no VL testing is available, while $\phi = 1$ means that it is available at all patient visits.

Tripartite diagnostic rules that satisfy (3) can be infinitely many, because if $\delta_{\mathcal{J}}(s)$ satisfies (3), so does $\delta_{\mathcal{J}'}(s)$ for all $\mathcal{J}' \subset \mathcal{J}$. We therefore restrict attention only to those rules that take maximum advantage of the available VL tests. All such rules form our decision space. Specifically, the decision space is defined as the set $\mathcal{D}_{\phi}^G = \{\delta_{\mathcal{J}} : G(u) - G(l) \leq \phi\}$ with the condition that for any $\delta_{\mathcal{J}}(s) \in \mathcal{D}_{\phi}^G$, there does not exist another rule $\delta_{\mathcal{J}'}(s)$ with $\mathcal{J}' \supset \mathcal{J}$ and satisfying (3).

For a given risk function $R(\cdot)$ and a decision space \mathcal{D}_{ϕ}^G , the optimal rule is defined as

$$\delta_{\mathcal{J}^*} = \arg \min_{\delta_{\mathcal{J}} \in \mathcal{D}_{\phi}^G} \{R(\mathcal{J})\}, \quad (4)$$

where \mathcal{J}^* indicates the optimal cut-offs on S for triaging the VL tests. We assume that the optimal rule is unique.

3.2 Optimal Rule Selection

In this section, we develop nonparametric and semiparametric approaches to determining the optimal rule from \mathcal{D}_{ϕ}^G . The nonparametric approach places no distributional assumption on either G_0 or G_1 and can therefore be broadly applied. The semiparametric approach assumes that G_0 and G_1 follow an exponential tilt model, whereby the densities $g_0(s)$ and $g_1(s)$ differ only by a factor proportional to $\exp(\beta_1 s)$, where β_1 is an unknown scalar parameter (called the tilting parameter). In Section 5, we use simulations to show that when the exponential tilt assumption holds, the semiparametric approach is generally more efficient in estimating the optimal rule when sample size is large.

Nonparametric Approach.

Suppose that we have a training data set of independent pairs $(S_1, Z_1), \dots, (S_n, Z_n)$. We first estimate G_1 , G_0 , and G empirically via

$$\begin{aligned} \hat{G}_z(s) &= \frac{\sum_{i=1}^n \mathbf{1}(S_i \leq s, Z_i = z)}{\sum_{i=1}^n \mathbf{1}(Z_i = z)}, \quad z = 0, 1, \\ \hat{G}(s) &= \hat{p}\hat{G}_1(s) + (1 - \hat{p})\hat{G}_0(s) \end{aligned}$$

with $\hat{p} = \sum_{i=1}^n Z_i/n$. To determine the optimal rule using (4), we then obtain the empirical decision space $\mathcal{D}_{\phi}^{\hat{G}}$ by the following steps:

1. Write $\tilde{\mathbf{L}} = (\tilde{l}_1, \tilde{l}_2, \dots, \tilde{l}_n)^{\top} = (S_{(1)}, S_{(2)}, \dots, S_{(n)})^{\top}$, where $\tilde{l}_j = S_{(j)}$ is the j -th order statistic of $\mathbf{S} = (S_1, \dots, S_n)^{\top}$.
2. For each \tilde{l}_j , calculate $\tilde{u}_j = \arg \max_{u \in \mathbf{S}} \{\hat{G}(u) - \hat{G}(\tilde{l}_j) \leq \phi\}$. Let $\tilde{\mathbf{U}} = (\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_n)^{\top}$.
3. For \tilde{u}_j and $\tilde{u}_{j'} \in \tilde{\mathbf{U}}$, $j < j'$, if $\tilde{u}_j = \tilde{u}_{j'}$, drop $\tilde{l}_{j'}$ from $\tilde{\mathbf{L}}$ and $\tilde{u}_{j'}$ from $\tilde{\mathbf{U}}$. Denote the resulting vectors by $\hat{\mathbf{L}} = (\hat{l}_1, \hat{l}_2, \dots, \hat{l}_m)^{\top}$ and $\hat{\mathbf{U}} = (\hat{u}_1, \hat{u}_2, \dots, \hat{u}_m)^{\top}$ with $m \leq n$.
4. The empirical decision space is given by $\mathcal{D}_{\phi}^{\hat{G}} = \{\delta_{\hat{\mathcal{J}}_j} : j = 1, 2, \dots, m\}$ with $\hat{\mathcal{J}}_j = [\hat{l}_j, \hat{u}_j]$.

With the empirical decision space $\mathcal{D}_\phi^{\widehat{G}}$, the optimal rule is then estimated via (4) with \mathcal{D}_ϕ^G replaced by $\mathcal{D}_\phi^{\widehat{G}}$. This can be carried out using a grid search. For example, to estimate the optimal rule that minimizes TMR, we calculate $\widehat{\text{FNR}}_j = \widehat{G}_1(\widehat{l}_j)$ and $\widehat{\text{FPR}}_j = 1 - \widehat{G}_0(\widehat{u}_j)$ for $j = 1, \dots, m$. Then, the optimal min-TMR rule is the rule in $\mathcal{D}_\phi^{\widehat{G}}$ that has a risk equal to $\min_j (\widehat{p}\widehat{\text{FNR}}_j + (1 - \widehat{p})\widehat{\text{FPR}}_j)$. Similarly, to identify the rule that minimizes $R_3(\mathcal{J}; \lambda)$ for a pre-specified λ , we select the rule in $\mathcal{D}_\phi^{\widehat{G}}$ that has a risk of $\min_j (\lambda\widehat{p}\widehat{\text{FNR}}_j + (1 - \lambda)(1 - \widehat{p})\widehat{\text{FPR}}_j)$.

Semiparametric Approach

The exponential tilt model has been used to characterize the relationship between components of a mixture distribution (Anderson 1972, 1979; Prentice and Pyke 1979; Efron 1981; Qin 1999). The model places no parametric assumptions on individual components of the mixture, except assuming that they differ only by a factor of the form

$$g_1(s) = \exp(\beta_0^* + \beta_1 s)g_0(s), \quad (5)$$

where β_1 is an unknown tilting parameter and $\beta_0^* = -\log E_{G_0}(e^{\beta_1 S})$ is a normalizing constant. Although no constraints are placed on g_0 , many commonly-used parametric distribution families can be represented in the form of (5), such as binomial, Poisson, normal with a common variance, and gamma distributions with a common shape parameter. In our case, the exponential tilt model is equivalent to the logistic model

$$\text{logit}\{\Pr(Z = 1 \mid S = s)\} = \beta_0 + \beta_1 s \quad (6)$$

with $\text{logit}(y) = \log\{y/(1 - y)\}$ and $\beta_0 = \beta_0^* + \text{logit}(p)$.

When the exponential tilt assumption holds, we can estimate G_0 and G_1 semiparametrically using the results in Appendix A.1, and then estimate the optimal rule using a grid-search in a similar way to what has been described in the last section.

If our goal is to identify a rule that minimizes TMR, it turns out that we can readily determine this rule without calculating the semiparametric estimates of G_0 and G_1 . To see this, we write $\Gamma(l, u, \tau) = R_1(\mathcal{J}) + \tau(G(u) - G(l) - \phi)$, and apply the Lagrange multiplier to solve $\partial\Gamma/\partial(l, u, \tau)^\top = 0$. It is straightforward to verify that the resulting rule must satisfy

$$l + u = -\frac{2\beta_0}{\beta_1}, \quad G(u) - G(l) \approx \phi.$$

That is, the optimal interval \mathcal{J}^* for triaging the limited VL testing is centered at $-\beta_0/\beta_1$, independent of the VL test availability ϕ . The optimal cut-off values therefore can be estimated by $\widehat{l} = -\widehat{\beta}_0/\widehat{\beta}_1 - \Delta_\phi$ and $\widehat{u} = -\widehat{\beta}_0/\widehat{\beta}_1 + \Delta_\phi$, where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are parameter estimates of the logistic model (6) and

$$\Delta_\phi = \arg \max_s \{\widehat{G}(-\widehat{\beta}_0/\widehat{\beta}_1 + s) - \widehat{G}(-\widehat{\beta}_0/\widehat{\beta}_1 - s) \leq \phi\}.$$

In the above equation, the empirical estimate \widehat{G} is used because the semiparametric estimate of G under the exponential tilt model is the same as \widehat{G} .

Uniqueness of Estimated Optimal Rule.

The estimated optimal rule based on a finite data set may not be unique, even though the true optimal rule is unique. When there are multiple rules that meet the optimality criterion, we propose to impose additional secondary criteria so as to determine a single optimal rule. For example, to determine an optimal rule from multiple rules that equally minimize $R_3(\mathcal{J}; \lambda)$, we may consider adding $R_1(\mathcal{J})$ as a secondary criterion and choosing one from these rules that has the lowest TMR. It is also reasonable to randomly choose one for practical use if the estimated optimal rules differ little.

4. ROC ANALYSIS

ROC analyses have been widely used to assess the overall diagnostic accuracy of bipartite classification rules. An ROC curve is a graphical presentation of the risk function $R_2(\cdot)$ associated with all candidate rules in a decision space. Comprehensive reviews of ROC analyses in biomedical research can be found in Pepe (2000), Zhou et al. (2002, Ch 2), Pepe (2003, Chs 4-5), and Gatsonis (2009). Recent applications of ROC analyses in studies of HIV-infected populations include Pahwa et al. (2008), Joska et al. (2011), and Mabeya et al. (2012), among many others.

4.1 ROC Curve for Tripartite Rules and AUC

ROC analyses for tripartite rules can be carried out in a fashion similar to conventional ROC analyses. With each rule in \mathcal{D}_ϕ^G represented by a point in a 2-dimensional space with its (FPR, $1 - \text{FNR}$) as the coordinates, an ROC curve for tripartite rules can be generated by connecting these points using a non-decreasing curve. Mathematically, we can express the ROC curve for tripartite rules as

$$C_\phi(t) : t \in [0, 1] \mapsto 1 - G_1 \circ H_\phi \circ G_0^{-1}(1 - t), \quad (7)$$

where $G_0^{-1}(t) = \inf\{s : G_0(s) \geq t\}$ is the generalized inverse of a cadlag function, $H_\phi(u) = \arg \inf_w \{G(u) - G(w) \leq \phi\}$, and ‘ \circ ’ denotes the function composition operator. Note that the difference between (7) and a conventional ROC curve is the operation induced by H_ϕ , which dictates that for each u and resulting FPR, the corresponding FNR is calculated based on a lower cut-off $l = H_\phi(u) \leq u$.

The area under the ROC curve (AUC) for tripartite rules is defined as,

$$\text{AUC}_\phi = \int_0^1 C_\phi(t) dt. \quad (8)$$

Like AUC for bipartite classification rules, AUC_ϕ provides an omnibus measure of diagnostic capability of all candidate rules in \mathcal{D}_ϕ^G . It can be interpreted as the true positive rate averaged across all FNRs. In Appendix A.2, we present several properties of AUC_ϕ , which turn out to be very similar to the AUC from a conventional ROC analysis.

As a special case when $\phi = 0$, $C_\phi(t)$ and AUC_ϕ reduce to conventional ROC curve and AUC for bipartite rules.

4.2 Estimation

With a training data set of independent $(S_1, Z_1), \dots, (S_n, Z_n)$ and a given ϕ , we can estimate the ROC curve for tripartite rules nonparametrically by

$$\widehat{C}_\phi(t) = 1 - \widehat{G}_1 \circ \widehat{H}_\phi \circ \widehat{G}_0^{-1}(1 - t), \quad (9)$$

where \widehat{G}_0 and \widehat{G}_1 are empirical estimates, and $\widehat{H}_\phi(u) = \arg \min_w \{\widehat{G}(u) - \widehat{G}(w) \leq \phi\}$. The estimated ROC curve is a step function with jumps only at points representing the rules in \mathcal{D}_ϕ^G . When the exponential tilt assumption holds, the ROC curve also can be estimated semiparametrically by replacing \widehat{G}_0 and \widehat{G}_1 in (9) by their corresponding semiparametric estimates given in Appendix A.1.

Using the results in the Appendix A.2 (See Eq.(A.1)), we can estimate AUC_ϕ nonparametrically by

$$\widehat{\text{AUC}}_\phi = \frac{1}{n^2 \widehat{p}(1 - \widehat{p})} \sum_{j=1}^n \sum_{i=1}^n Z_i(1 - Z_j) \left[\mathbf{1}\{S_i > \widehat{H}_\phi(S_j)\} + \frac{\mathbf{1}\{S_i = \widehat{H}_\phi(S_j)\}}{2} \right]. \quad (10)$$

In Appendix A.3, we present several large-sample properties of the nonparametric estimates $\widehat{C}_\phi(t)$ and $\widehat{\text{AUC}}_\phi$.

4.3 Using ROC curve for Rule Selection

An ROC curve for tripartite rules also can be used for optimal rule selection, recognizing that the diagnostic properties of each rule in \mathcal{D}_ϕ^G are characterized by a point on the curve. For example, if the ROC curve is smooth and concave, it can be verified that the min- λ rule corresponds to the point on the ROC curve where the tangent is equal to $(1 - \lambda)(1 - p)/(\lambda p)$ (Metz 1978). Broader discussions on using ROC curves for optimal rule selection can be found in Zhou et al. (2002) and Pepe (2003).

5. SIMULATION STUDIES

In this section, we conduct simulation studies to examine 1) the diagnostic accuracy of the optimal rules estimated by the nonparametric and semiparametric approaches, and 2) the large-sample properties of the estimated optimal rules. For the first aim, we consider two scenarios when the exponential tilt assumption is and is not satisfied. For the second aim, we focus on the setting where the exponential tilt assumption holds. For simplicity, we consider estimating the optimal rules that minimize TMR.

We use the negative value of CD4 count as a risk score. We first simulate viral status Z assuming that $Z \sim \text{Bernoulli}(p)$, and then conditional on Z , simulate $(\text{CD4}|Z = z) = \lceil W \rceil$ with $W \sim \text{Gamma}(\eta_z, \kappa_z)$, where $\lceil \cdot \rceil$ denotes the ceiling operation, and η_z and κ_z are the shape and scale parameters of the gamma distribution.

Scenario (A) considers the case when the exponential tilt assumption does not hold. We conduct two simulations by simulating CD4 count data from gamma distributions with parameters,

$$(A-1): \quad (\eta_0, \kappa_0) = (3.2, 152) \text{ and } (\eta_1, \kappa_1) = (2.3, 133)$$

$$(A-2): \quad (\eta_0, \kappa_0) = (4.8, 100) \text{ and } (\eta_1, \kappa_1) = (2.3, 133).$$

The parameter values of (A-1) are chosen as the maximum likelihood estimates (MLEs) obtained by fitting gamma distributions to the Miriam Immunology Clinic data (which will be analyzed in Section 6). For (A-2), we choose the same values of (η_1, κ_1) as in (A-1), but set (η_0, κ_0) such that the exponential tilt assumption is further violated while keeping $\eta_0 \kappa_0$ unchanged, i.e. the average CD4 count for patients without treatment failure stays the same as (A-1). (Recall that the mean of gamma distribution is $\eta \kappa$.)

Scenario (B) considers the case when the exponential tilt assumption holds. We choose a common shape parameter $\eta_0 = \eta_1 = 2.8$, the mid-point of η_0 and η_1 in (A-1), and conduct two simulations with parameters

$$\begin{aligned} \text{(B-1): } & (\eta_0, \kappa_0) = (2.8, 173) \text{ and } (\eta_1, \kappa_1) = (2.8, 111) \\ \text{(B-2): } & (\eta_0, \kappa_0) = (2.8, 350) \text{ and } (\eta_1, \kappa_1) = (2.8, 111). \end{aligned}$$

The values of κ_0 and κ_1 in (B-1) are the MLEs obtained by fitting gamma distributions to the Miriam Immunology Clinic data while fixing their shape parameters at 2.8. For (B-2), we choose a large value of $\kappa_0 = 350$ to simulate a case when two gamma distributions are further separated. The gamma densities of the four simulations are shown in Figure 2.

5.1 Diagnostic Accuracy

For the first aim, we consider three prevalences of treatment failure, $p = (.15, .25, .40)$, and assume that VL testing is available at proportions $\phi = (0, 20, 40)\%$ of patient visits. For each parameter combination, we simulate 1000 datasets each having 5000 observations. The first 2500 observations of each dataset are used as the training data to develop an optimal rule, and the remaining 2500 observations as the testing data to calculate its associated misclassification rate. Results are shown in Table 1.

When the prevalence of treatment failure is low (e.g., $p \leq .25$) and VL test availability ϕ is high, the semiparametric approach may yield a negative estimate of the lower cut-off value on CD4 count, particularly when the center of the optimal cut-off interval is close to zero. When this occurs, we replace the negative cut-off values by zero. This adjustment does not imply that the algorithm fails. It can be verified that the upper cut-off estimate is still correct, and the optimal rule in this case is to assign VL test to those high-risk patients with CD4 count less than the upper cut-off value.

Table 1 shows that the nonparametric approach yields the correct estimates of the optimal cut-off values for both Scenarios (A) and (B), and the resulting TMRs are close to the underlying truth. When the exponential tilt assumption does not hold as in Scenario (A), the optimal rules estimated by the semiparametric approach are slightly biased (contrasted with Scenario (B)). However when the exponential tilt assumption holds as in Scenario (B), the semiparametric approach yields the correct estimates of the optimal cut-off values, and the estimated cut-off values have much smaller standard errors compared with their corresponding nonparametric estimates.

5.2 Convergence Rate and Efficiency

The second aim of our simulation studies is to examine the relative efficiency of the nonparametric and semiparametric approaches. For this aim, we consider only the parameter setup

Figure 2: Gamma distributions used for simulating CD4 count data. The gray step lines in the top-left subplot are histograms of the CD4 data from the Miriam Hospital Immunology Clinic. The smooth dashed (solid) lines are gamma densities for those with (without) treatment failure.

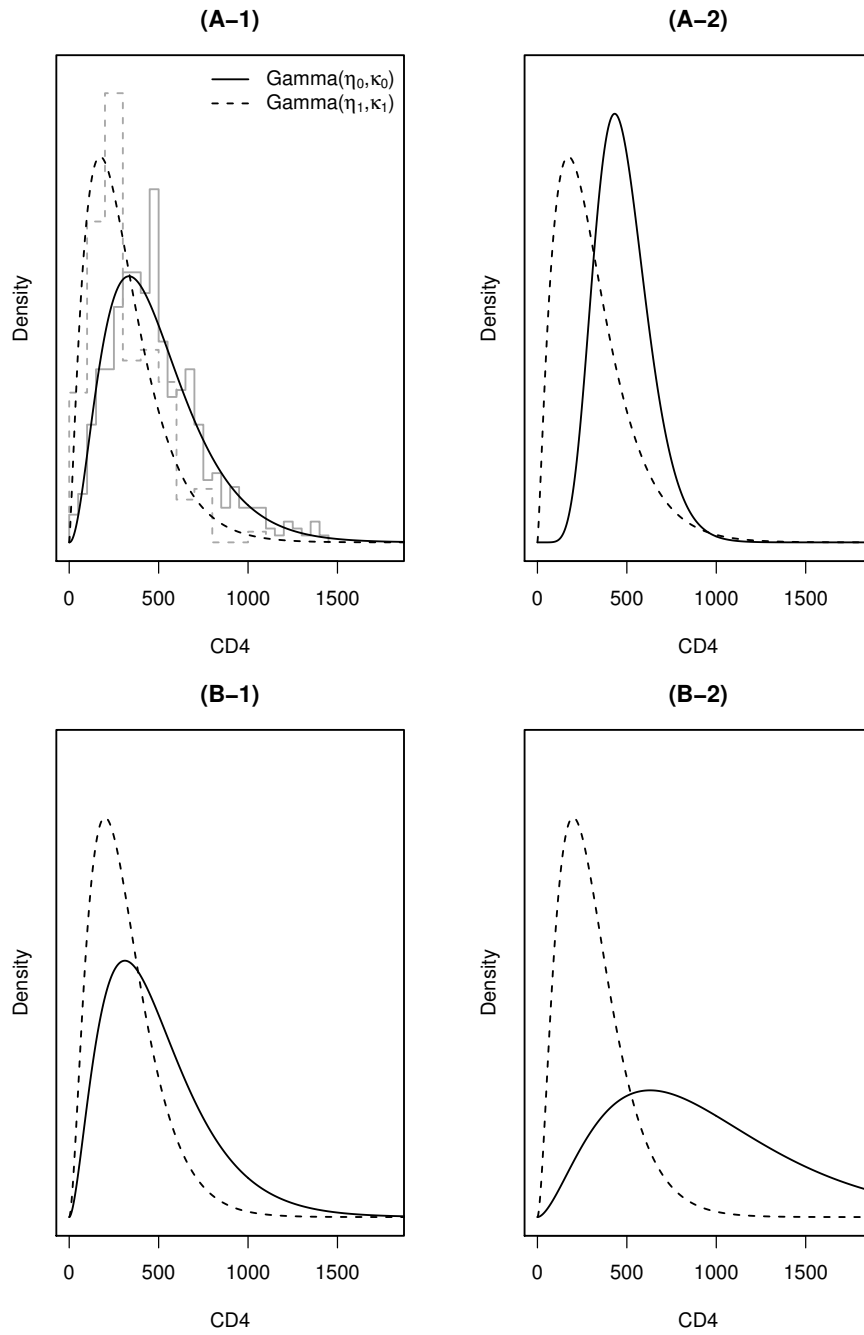
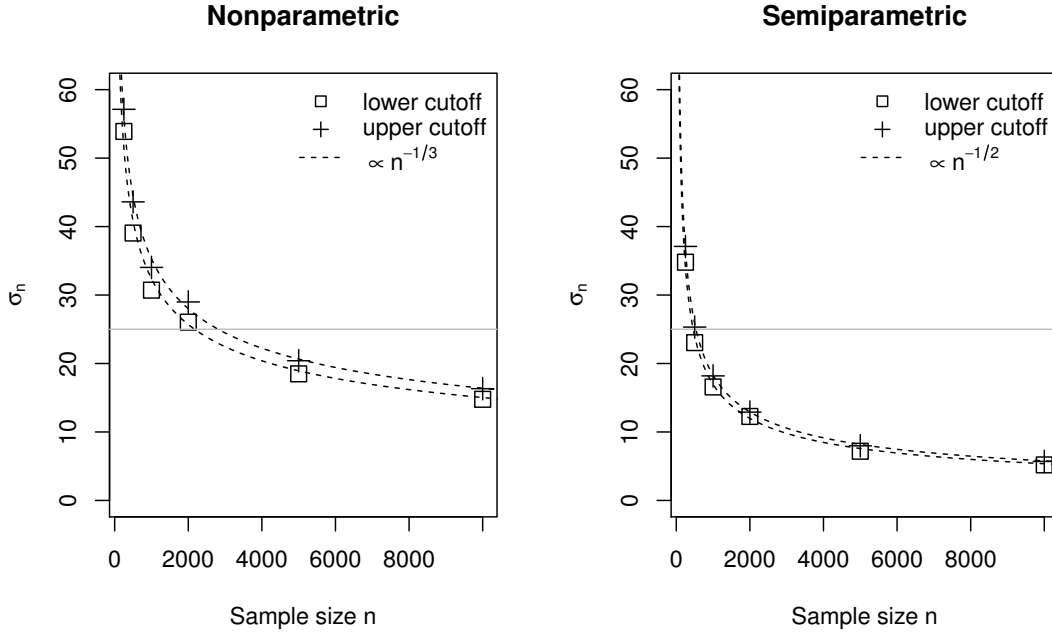


Table 1: Simulation results. For each condition, the estimated lower and upper cut-off values are averages over 1000 simulations, and converted to the original scale of CD4 count. The numbers in parentheses are standard errors.

p	ϕ	true values			nonparametric estimate			semiparametric estimate		
		lower	upper	TMR	lower	upper	TMR	lower	upper	TMR
(A-1)	.15	0	65	.15	70 (20)	70 (20)	.148	0 (1)	0 (1)	.150
		.2	0	.09	29 (14)	231 (5)	.086	0 (0)	230 (5)	.086
		.4	0	.05	25 (13)	349 (6)	.051	0 (0)	348 (6)	.051
	.25	0	125	.24	128 (23)	128 (23)	.237	79 (25)	79 (25)	.240
		.2	66	.15	71 (21)	228 (9)	.152	2 (8)	214 (5)	.153
		.4	39	.09	47 (20)	336 (7)	.092	0 (0)	329 (6)	.092
	.4	0	239	.32	240 (28)	240 (28)	.327	268 (13)	268 (13)	.327
		.2	177	.23	192 (22)	302 (22)	.233	212 (13)	323 (14)	.234
		.4	149	.15	148 (22)	377 (22)	.153	153 (14)	382 (14)	.152
	(A-2) .15	0	130	.14	132 (16)	132 (16)	.135	35 (23)	35 (23)	.147
		.2	72	.07	71 (20)	268 (6)	.075	0 (0)	262 (5)	.075
		.4	57	.04	59 (21)	374 (6)	.045	0 (0)	370 (5)	.045
(B-1)	.25	0	176	.21	178 (17)	178 (17)	.209	148 (17)	148 (17)	.212
		.2	122	.13	121 (17)	269 (9)	.131	51 (28)	247 (7)	.135
		.4	98	.08	96 (18)	368 (8)	.080	1 (3)	351 (5)	.083
	.4	0	249	.29	251 (20)	251 (20)	.291	290 (10)	290 (10)	.295
		.2	200	.20	203 (17)	314 (15)	.201	236 (11)	345 (11)	.204
		.4	167	.13	169 (18)	396 (15)	.130	177 (11)	403 (10)	.130
	(B-2) .15	0	0	.15	27 (15)	27 (15)	.150	0 (0)	0 (0)	.150
		.2	0	.10	20 (9)	221 (5)	.095	0 (0)	221 (5)	.095
		.4	0	.05	18 (8)	338 (6)	.056	0 (0)	338 (6)	.055
	.25	0	45	.25	69 (34)	69 (34)	.251	45 (25)	45 (25)	.250
		.2	0	.17	33 (19)	211 (6)	.166	0 (2)	209 (5)	.165
		.4	0	.10	26 (14)	323 (6)	.100	0 (0)	322 (6)	.099
	.4	0	259	.35	257 (34)	257 (34)	.355	259 (15)	259 (15)	.353
		.2	215	.26	206 (28)	312 (29)	.259	207 (15)	312 (15)	.258
		.4	159	.17	148 (31)	370 (30)	.171	149 (16)	369 (15)	.170
	(B-2) .15	0	241	.13	241 (31)	241 (31)	.127	242 (13)	242 (13)	.126
		.2	99	.05	103 (37)	391 (19)	.047	97 (20)	386 (11)	.047
		.4	0	.01	38 (19)	622 (14)	.011	0 (0)	620 (13)	.010
(B-2)	.25	0	344	.16	345 (30)	345 (30)	.165	345 (10)	345 (10)	.164
		.2	234	.08	237 (22)	455 (25)	.082	236 (10)	454 (11)	.081
		.4	112	.03	111 (33)	581 (26)	.028	111 (13)	578 (11)	.028
	.4	0	457	.18	458 (30)	458 (30)	.184	458 (9)	458 (9)	.183
		.2	344	.10	347 (20)	567 (30)	.100	347 (8)	568 (12)	.100
		.4	237	.05	240 (19)	677 (38)	.047	239 (8)	676 (13)	.046

Figure 3: Large-sample convergence properties of estimated optimal cut-off boundaries. Horizontal lines are added to indicate the sample sizes needed to achieve $\sigma_n = 25$.



of (B-2) with $p = .25$ and $\phi = .20$, but simulate the training data with increasing sample sizes of $n = (250, 500, 1000, 2000, 5000, 10000)$. For each sample size, we simulate 1000 training datasets, and for each data set, we estimate the optimal rules both nonparametrically and semiparametrically.

With a slight abuse of notation, we use σ_n^2 to denote the variances of both estimated upper and lower cut-off values. Then assuming that $\sigma_n \propto n^{-w}$ (a sufficient condition for $\sigma_n = O(n^{-w})$ as $n \rightarrow \infty$), we use simulations to approximate w for the two estimation approaches. Specifically, we compute σ_n based on the 1000 estimated optimal cut-off values for each sample size n . We then regress $\log(\sigma_n)$ on $(-\log n)$ using a simple linear model with a slope w . By least-squares estimation, we find that $w \approx 0.33$ when the optimal cut-off values are estimated nonparametrically, and ≈ 0.50 when they are estimated semiparametrically. The results are shown in Figure 3.

The simulations suggest that in this specific case, the nonparametric estimates of the optimal cut-off values converge approximately at a rate of $O(n^{-1/3})$, and the semiparametric estimates converge at a faster rate of about $O(n^{-1/2})$. The relative efficiency between the two estimation approaches is approximately $O(n^{1/6})$ when n is large.

5.3 Simulation-Based Study Design

The results above also suggest that a study for tripartite rule development can be designed based on simulations. For example, suppose that the same assumptions as in Section 5.2 are made, and we would like to design a study to determine an optimal tripartite rule such that the 95% confidence intervals of both upper and lower cut-offs have widths of no more than

Table 2: Summary statistics for key variables ($n = 597$)

Marker	mean	median	IQR	range
<i>Virological marker</i>				
VL at most recent visit (copies/mL)	11.8K	75	(75, 400)	(12, > 500K)
<i>Immunological markers</i>				
CD4 count at most recent visit (cells/uL)	442	407	(254, 576)	(8, 1412)
6-month CD4 count change (%)	7.3	18	(−13, 33)	(−80, 736)
CD4 % at most recent visit	24	23	(17, 30)	(.90, 59)
6-month CD4% change (%)	9.5	4.7	(−6.1, 16)	(−74, 209)

K: thousand; IQR: Interquartile range.

100 CD4 (i.e., $\sigma_n \leq 25$). Then referring to the gray horizontal lines in Figure 3, a study with a sample size of about 3000 subjects is needed if the nonparametric approach is used to estimate the optimal rule, or a sample size of about 500 if the exponential tilt assumption holds and the semiparametric approach is used.

6. APPLICATION

6.1 Data from the Miriam Hospital Immunology Clinic

For illustration, we analyze data from the Miriam Immunology Clinic in Providence, RI, USA, the largest HIV clinic in the state (Gillani 2009). We recognize the essential difference between HIV-infected patients in the US and RLS. The main reason we use a US dataset to demonstrate the development of clinical rules is because this database contains high quality CD4 and VL data that were routinely and frequently collected.

We use data from the most recent clinic visits of 597 patients who meet the following criteria: have taken ART for at least 6 months; have CD4 count, CD4% and VL measure available at the most recent clinic visit; and have CD4 count and CD4% available 6 months (with a window of 6 ± 1 mo) prior to that visit. We calculate the 6-month changes in CD4 count and in CD4%, where [6-month change] is defined as $([\text{current}] - [\text{6-mo ago}]) / [\text{6-mo ago}]$. Total time on ART, while a potentially important predictor (Kantor et al. 2009), is not available for all patients and therefore not used in formulating risk scores.

Table 2 provides summary statistics for key clinical and immunological markers in the data. For uniformity, viral failure is defined as having VL above 400 copies/mL (some of the VL test assays used have lower detection limits of < 400 copies/mL). Among the 597 patients, 146 have viral failure; so the estimated prevalence of viral failure is $\hat{p} = 146/597 = .25$.

6.2 Risk Scores

Two risk scores are considered for developing diagnostic rules. The first risk score is $S_1 = -\text{CD4}$, or negative value of the most recent CD4 count (to be consistent with the notion that greater values of S correlate with increased risk of viral failure). The second risk score is a prediction-based composite score derived from a logistic regression of treatment failure

Table 3: Diagnostic accuracies of the two simple empirical rules.

Diagnosis action based on CD4						
ϕ	test positive	request VL test	test negative	FPR	FNR	TMR
0	0 - 200	–	≥ 200	.10	.70	.26
.15	–	0 - 200	≥ 200	0	.70	.18

on four immunological markers as follows,

$$S_2 = \text{logit}^{-1}(\beta_0 + \beta_1[\text{CD4}] + \beta_2[\text{CD4 \%}] + \beta_3[6\text{-mo CD4 change}] + \beta_4[6\text{-mo CD4\% change}]),$$

where CD4 and CD4% refer to their measures at current visit. The MLEs (SEs) of the coefficients are $\hat{\beta}_0 = .89$ (.27), $\hat{\beta}_1 = -.0021$ (.00074), $\hat{\beta}_2 = -.049$ (.017), $\hat{\beta}_3 = -.055$ (.21), and $\hat{\beta}_4 = -1.40$ (.46). A Hosmer-Lemeshow test gives a p-value of .28, indicating no evidence of lack of fit. The distribution of S_2 has a median .21, ranges from .01 to .87, and can be interpreted as the predicted probability of treatment failure.

The risk score S_1 is easier to implement in clinical practice but known to have a high error rate for diagnosing viral failure. By incorporating more clinical information, S_2 is potentially more accurate, but its use in clinical settings is not as straightforward as S_1 .

6.3 Two Simple Rules

Before calculating tripartite rules based on criteria laid out in Section 3, we summarize operating characteristics of two simple diagnostic rules that are similar in spirit to those commonly used in RLS when VL test has limited or no availability.

The first rule assumes that no VL testing is available (i.e., $\phi = 0$) and uses $\text{CD4} < 200$ as the hard cut-off for diagnosing treatment failure and $\text{CD4} \geq 200$ as non-failing, a criterion recommended by the WHO for the RLS (WHO 2010b). (Another criterion recommended by the WHO for the RLS is using $\text{CD4} = 350$ as the cut-off threshold.)

The second rule assumes that the limited VL testing will be used only as a confirmative test for patients with $\text{CD4} < 200$. This rule classifies those with CD4 count ≥ 200 as non-failure, and makes correct diagnoses for patients with CD4 count < 200 . In the Miriam Immunology Clinic data, about 15% of patients have current CD4 count less than 200, so we consider the case that VL testing is available at 15% of patient visits, i.e., $\phi = .15$.

The diagnostic accuracies of these two rules are summarized in Table 3. Both rules have FNR around 0.70. The second rule, by having 15% of patients tested for VL, reduces the FPR to 0 and TMR from .26 to .18. The improvement realized by having VL testing available to a small fraction of patients is evident; however, whether the second rule is optimal needs further investigations.

6.4 Analysis I: CD4-Based Min- λ Rules

In this section, we evaluate the diagnostic performance of optimal tripartite rules based on S_1 , using $R_3(\cdot)$ as the risk criterion. To make a direct comparison to the simple rules in the last section, we assume that VL testing is available at 15% of patient visits. The optimal tripartite rules are developed using the nonparametric approach as described in Section 3.

Figure 4: The optimal min- λ rules based on S_1 and associated FPR and FNR.

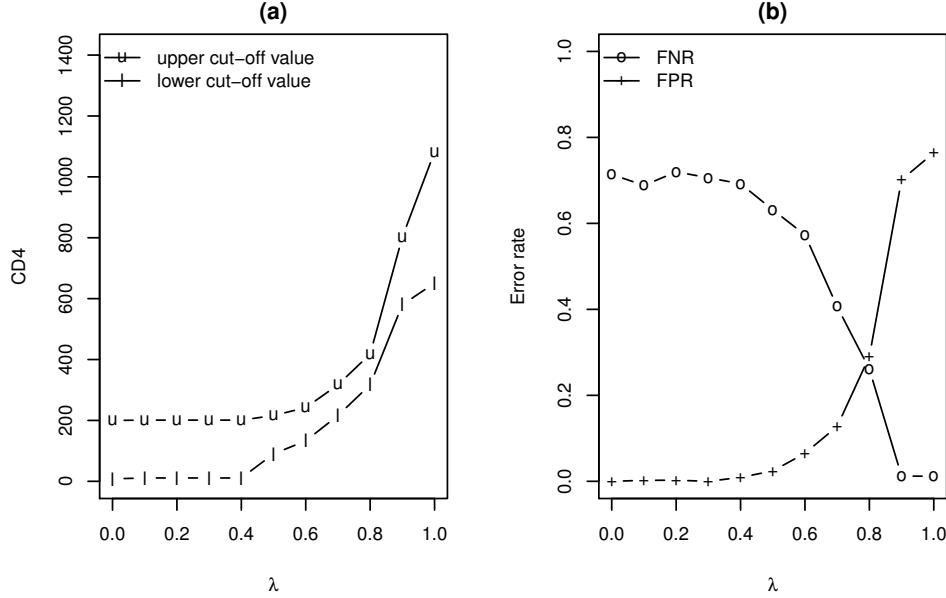


Figure 4 shows the estimated optimal rules and associated FNR and FPR, for λ varying from 0 to 1. The FNR and FPR are computed using 10-fold cross validations, carried out as follows. We randomly subdivide the data into 10 subsets of about equal size; determine the FPR and FNR for each subset using the optimal rule developed using the remaining 9 subsets; and then calculate the FPR and FNR as the averages over the 10 pieces (cf. Hastie et al. 2001).

As shown in Figure 4, when λ increases (i.e., placing higher priority on avoiding false negative diagnoses), the estimated optimal rule shifts gradually toward triaging the VL tests to those with CD4 count in the middle and high range. At $\lambda = .8$, the estimated optimal rule calls for testing patients having CD4 count between 300 and 450; in this case, both FNR and FPR are around .30. At the extreme when $\lambda = 1$, the estimated optimal rule calls for VL tests on those with $CD4 > 650$, which reduces the FNR to ≈ 0 but increases FPR to $\approx .80$.

The left panel of Figure 4 shows that when $\lambda < .4$, the estimated optimal rule is to obtain VL when $17 < CD4 < 201$. That is, when avoidance of false positive diagnosis is prioritized, the simple rule using VL testing as a confirmative test is optimal and a reasonable choice.

6.5 Analysis II: Comparing S_1 - and S_2 -Based Rules that Minimize the Weighted Risk R_3
Next, we compare the diagnostic accuracy of single-marker tripartite rules based on S_1 to multiple-marker rules based on S_2 using $R_3(\cdot)$ as the risk criterion. We consider three values of $\lambda = (.25, .50, .75)$ and three constraints on VL test availability $\phi = (0, .15, .30)$. Nonparametric estimates of the optimal rules, along with FPR, FNR and TMR obtained from cross-validations, are given in Table 4. Standard errors for all table entries are computed using the bootstrap method with 500 re-samples.

In summary, the optimal rules based on S_2 have a slightly better diagnostic performance

Table 4: Comparison of the S_1 - and S_2 -based tripartite rules. The optimal cut-off points based on S_1 are transformed back to the original scale of CD4 count. The numbers in parentheses are standard errors.

	λ	ϕ	cut-off points		FNR	FPR	R_3	TMR
			lower	upper				
S_1	.25	.00	17 (17)	17 (17)	.98 (.02)	.00 (.00)	.06 (.01)	.24 (.02)
		.15	17 (9)	201 (18)	.72 (.04)	.00 (.00)	.04 (.00)	.18 (.01)
		.30	17 (13)	284 (13)	.46 (.04)	.00 (.00)	.03 (.00)	.12 (.01)
	.50	.00	120 (64)	120 (64)	.93 (.12)	.03 (.04)	.12 (.01)	.26 (.02)
		.15	90 (16)	216 (39)	.63 (.06)	.02 (.01)	.08 (.01)	.17 (.01)
		.30	17 (15)	284 (32)	.45 (.04)	.01 (.01)	.05 (.01)	.12 (.01)
	.75	.00	302 (45)	302 (45)	.43 (.06)	.26 (.06)	.13 (.01)	.30 (.04)
		.15	216 (51)	317 (50)	.40 (.07)	.13 (.06)	.10 (.01)	.20 (.04)
		.30	226 (67)	417 (81)	.30 (.07)	.14 (.06)	.08 (.01)	.18 (.04)
S_2	.25	.00	.64 (.04)	.64 (.04)	.91 (.04)	.01 (.00)	.06 (.00)	.23 (.01)
		.15	.39 (.01)	.75 (.04)	.62 (.04)	.00 (.00)	.04 (.00)	.16 (.01)
		.30	.29 (.01)	.71 (.04)	.42 (.04)	.00 (.00)	.03 (.00)	.11 (.01)
	.50	.00	.53 (.07)	.53 (.07)	.79 (.08)	.04 (.02)	.11 (.01)	.23 (.01)
		.15	.37 (.01)	.66 (.06)	.60 (.05)	.01 (.01)	.07 (.01)	.16 (.01)
		.30	.28 (.01)	.67 (.06)	.43 (.04)	.01 (.01)	.05 (.01)	.11 (.01)
	.75	.00	.26 (.04)	.26 (.04)	.35 (.08)	.30 (.08)	.12 (.01)	.31 (.04)
		.15	.26 (.04)	.34 (.05)	.32 (.08)	.21 (.07)	.09 (.01)	.24 (.03)
		.30	.19 (.03)	.34 (.08)	.25 (.06)	.13 (.05)	.07 (.01)	.16 (.03)

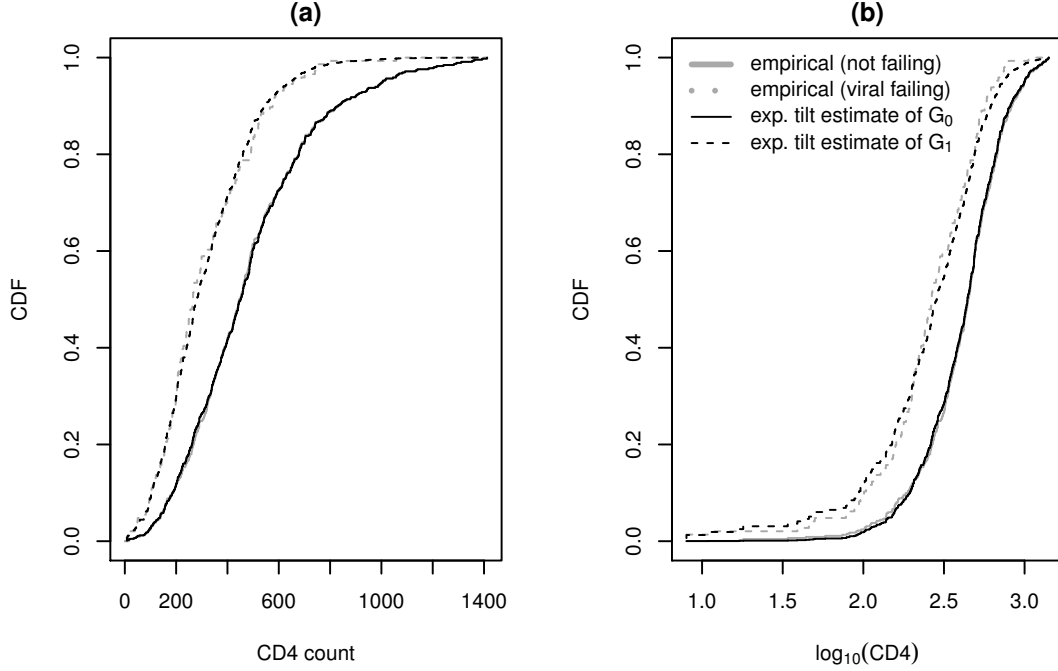
than the optimal rules on S_1 . However, the magnitude of improvement by incorporating more non-VL markers is small, relative to the improvement that can be achieved by the selective use of VL testing on more patients. In Section 6.7, the diagnostic accuracies of the rules based on the two risk scores will be further compared using AUCs.

6.6 Analysis III: Optimal Rules under Exponential Tilt Assumption

In this section, we develop the optimal tripartite rules under the exponential tilt assumption. We consider the following two risk scores, S_1 and $S_1^* = -\log_{10}(\text{CD4})$, for rule development. The reason for using S_1^* is that it avoids the issue of having the empirical adjustment when one cut-off value is beyond the support of the risk score (as we encountered in our simulation studies). The risk score S_1^* may also be more suitable for the exponential tilt model.

We first examine the suitability of the exponential tilt model for S_1 and S_1^* by plotting the semiparametric estimates of G_0 and G_1 against their empirical estimates. The results are shown in Figure 5, where the semiparametric estimates of G_0 and G_1 are obtained using the results in Appendix A.1. Figure 5 suggests that the exponential tilt assumption is reasonable for both S_1 and S_1^* although the goodness of fit for S_1 is slightly better. (One also can use

Figure 5: Empirical and semiparametric estimates of the cumulative densities of CD4 counts and $\log_{10}(\text{CD4})$.



Q-Q plots, not shown, to examine the model goodness of fit.)

The estimated optimal rules using TMR as the risk criterion are given in Table 5. The intervals for triaging VL assays are centered at $\text{CD4} = 77$ and 109 for the optimal rules based on S_1 and S_1^* , respectively. Overall, the diagnostic accuracies of the two sets of estimated optimal rules are comparable, and their estimated cut-off values differ only slightly relative to their standard errors. The optimal rules in Table 5 also are comparable to the optimal rules that are developed nonparametrically (see Table 4 with $\lambda = .50$), but in general have much smaller standard errors.

6.7 Analysis IV: ROC Analyses for Tripartite Rules

Figure 6 shows nonparametric estimates of ROC curves for tripartite rules based on S_1 and S_2 , when the VL tests are available at $\phi = 0, 15, 30, 45$, and 60% of patient visits. The ROC curves in the subplot (b) are slightly better than those in the subplot (a), which suggests that improvement in diagnostic capacity can be achieved using the composite score S_2 . See also the AUC curves and their difference in the subplots (c) and (d). Consistent with our findings in Analysis II (Section 6.5), the difference between the two AUC curves, although statistically significant for $\phi < .6$, is marginal.

Relative to not having VL tests available, the AUCs for tripartite rules based on both risk scores are substantially improved as VL testing is made available for some of clinical visits. For example, as shown in the subplot (c), when we increase the VL test availability from 0 to 20%, the absolute improvement in AUC is about 15%; and increasing availability to 40% improves AUC by more than 20%. In particular, the *relative* improvement by making VL

Table 5: The optimal min-TMR rules under the exponential tilt assumption. Cut-off points were transformed back to the original scale of CD4 count. The numbers in parentheses are standard errors calculated using the bootstrap method.

		cut-off points		FNR	FPR	TMR
	ϕ	lower	upper			
S_1	0	77 (46)	77 (46)	.94 (.06)	.01 (.02)	.24 (.02)
	.15	0 (29)	199 (10)	.69 (.05)	.00 (.01)	.17 (.02)
	.30	0 (10)	278 (11)	.47 (.04)	.00 (.00)	.11 (.01)
S_1^*	0	109 (23)	109 (23)	.89 (.05)	.03 (.01)	.24 (.02)
	.15	57 (22)	206 (10)	.67 (.04)	.01 (.01)	.18 (.02)
	.30	41 (17)	287 (12)	.46 (.04)	.01 (.00)	.12 (.01)

testing accessible to some HIV patients is more pronounced when the VL testing availability is low.

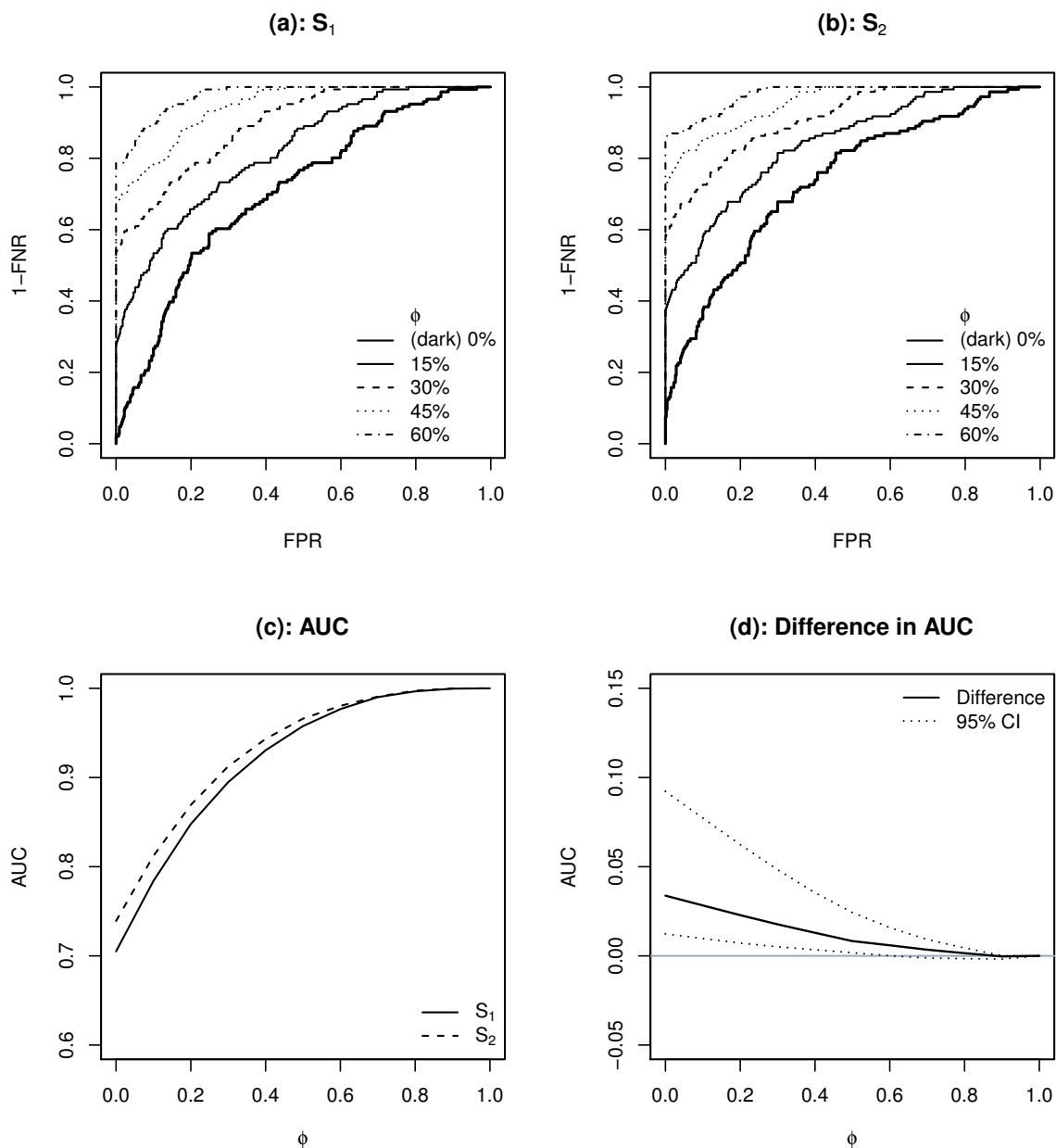
7. DISCUSSION

This paper is motivated by recent evidence that the CD4-based WHO guidelines for monitoring HIV treatment in RLS can lead to high treatment failure misclassification rates, and by the fact that VL tests are becoming available to programs and patients in RLS, typically on a limited basis. To make optimal use of VL tests, we propose a tripartite diagnostic rule based on a risk score that subdivides patients into a high-risk group (classified as treatment failure), a low-risk group (as viral suppressed), and an intermediate-risk group to whom the limited VL tests are assigned, where the size of the third group is constrained by the availability of VL tests. Nonparametric and semiparametric methods are proposed for determining an optimal rule to minimize a given risk criterion. ROC analysis procedure for characterizing the diagnostic performance of tripartite rules and its associated asymptotic properties are developed.

Our proposed method is demonstrated by analyzing data from the Miriam Hospital Immunology Clinic in Providence, RI. We show that with selective and targeted use of VL tests, the rate of misdiagnosis can be substantially reduced even when VL testing is available at a small portion of patient visits (e.g., $\phi = 15\%$). Our analysis also suggests that when avoidance of false positive diagnoses is prioritized, using VL testing strictly to confirm viral failure for those deemed to be at high risk is a reasonable choice. This finding applies only to patients at the Miriam Hospital Immunology Clinic; its external validity remains to be tested.

Our methods assume that the functional form of risk score $S(\mathbf{X})$ is known, but may be relaxed by unknown parameters. When the function form of $S(\mathbf{X})$ is unknown, methods of machine/statistical learning, such as boosting (Freund et al. 1999), targeted/super learner (Sinisi et al. 2007; van der Laan 2011), classification tree learning (Breiman et al. 1984), neural networks (Hagan et al. 1996; Sarle 1994), and prediction-based classification methods (Foulkes and De Gruttola 2002), can be implemented. We refer the readers to Hastie et al.

Figure 6: ROC curves for diagnostic rules using S_1 and S_2 (subplots (a) and (b)); the resulting AUC curves as functions of ϕ (subplot (c)); and the difference of the two AUC curves (S_2 “minus” S_1 , subplot (d)). The point-wise 95% CI of the difference in AUC is calculated using the bootstrap method with 500 re-samples.



(2001) and Kotsiantis (2007) for a more comprehensive treatment on the topic.

We assume that there is no measurement error in VL, i.e. that VL test is the gold standard for determining the amount of circulating virus. In developed countries, repeated VL tests and HIV genotyping are usually required to confirm treatment failure and existence of drug resistance once the VL becomes detectable. HIV-infected patients in RLS however do not have such luxury, and a single VL test result (if done) is probably the most direct measure of viral failure, and is used for clinical decision making. So although the assumption is not ideal, it is reasonable in this context because the measured VL is the best available basis for decision making in RLS. Future work will address the issue of measurement error in VL and its effect on misclassification rates.

CD4 counts are known to be highly variable due to measurement error, diurnal variations, and other factors. The measurement error of CD4 count may be part of the reason for high misclassification rates of the WHO guidelines. The impact of measurement error in biomarkers on predicting binary outcomes has been studied by Carroll et al. (1984, 2006), Buzas et al. (2003), and Fuller (2009) among others. Generally speaking, large measurement errors of a biomarker are associated with a greater attenuation of its capability of predicting outcomes. One way to reduce the impact of measurement errors is through repeated measurements. Given the fact that point-of-care CD4 technologies are being developed, it may be possible in practice in the future to quantify and reduce the impact of CD4 measurement error by multiple testing at a single visit. On the other hand, with additional information such as prior history of CD counts, it may also be possible to evaluate the magnitude of measurement error by constructing appropriate measurement error models (which typically rely on certain subjective assumptions) and applying methods such as regression calibration (Carroll and Stefanski 1990; Rosner et al. 1990) and simulation-extrapolation (Stefanski and Cook 1995). Improving diagnostic accuracy by reducing the impact of CD4 measurement error is an area worthwhile further investigation.

A final limitation of this paper, as it applies to developing rules for RLS, is that a US data set is used to demonstrate our proposed methods. Our ongoing work is focused on developing and calibrating rules based on data from sub-Saharan Africa and other RLS.

REFERENCES

- Abouyannis, M., Menten, J., Kiragga, A., Lynen, L., Robertson, G., Castelnovo, B., Manabe, Y. C., Reynolds, S. J., and Roberts, L. (2011), “Development and validation of systems for rational use of viral load testing in adults receiving first-line ART in sub-Saharan Africa,” *AIDS*, 25, 1627–1635.
- Anderson, A. M. and Bartlett, J. A. (2006), “Changing antiretroviral therapy in the setting of virologic relapse: review of the current literature,” *Current HIV/AIDS Reports*, 3, 79–85.
- Anderson, J. A. (1972), “Separate sample logistic discrimination,” *Biometrika*, 59, 19–35.
- (1979), “Multivariate Logistic Compounds,” *Biometrika*, 66, 17–26.
- Bagchi, S., Kempf, M., Westfall, A., Maherya, A., Willig, J., and Saag, M. (2007), “Can routine clinical markers be used longitudinally to monitor antiretroviral therapy success in resourcelimited settings?” *Clinical Infectious Diseases*, 44, 135–138.

- Baker, S. (2009), “Putting risk prediction in perspective: relative utility curves,” *Journal of the National Cancer Institute*, 101, 1538–1542.
- Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag, 2nd ed.
- Bisson, G. P., Gross, R., Bellamy, S., Chittams, J., Hislop, M., Regensberg, L., Frank, I., Maartens, G., and Nachega, J. B. (2008), “Pharmacy refill adherence compared with CD4 count changes for monitoring HIV-infected adults on antiretroviral therapy,” *PLoS Medicine*, 5, e109.
- Bisson, G. P., Gross, R., Strom, J. B., Rollins, C., Bellamy, S., Weinstein, R., Friedman, H., Dickinson, D., Frank, I., Strom, B. L., Gaolathe, T., and Ndwapu, N. (2006), “Diagnostic accuracy of CD4 cell count increase for virologic response after initiating highly active antiretroviral therapy,” *AIDS*, 20, 1613–1619.
- Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984), *Classification and regression trees*, Chapman & Hall/CRC.
- Buzas, J., Tosteson, T., and Stefanski, L. (2003), “Measurement error,” *Institute of Statistics Mimeo Series*, No.2544, 4, 1–92.
- Calmy, A., Ford, N., Hirschel, B., Reynolds, S. J., Lynen, L., Goemaere, E., de la Vega, F. G., Perrin, L., and Rodriguez, W. (2007), “HIV viral load monitoring in resource-limited regions: optional or necessary?” *Clinical Infectious Diseases*, 44, 128–134.
- Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006), *Measurement error in nonlinear models: a modern perspective*, vol. 105, Chapman & Hall/CRC.
- Carroll, R. and Stefanski, L. (1990), “Approximate quasi-likelihood estimation in models with surrogate predictors,” *Journal of the American Statistical Association*, 85, 652–663.
- Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., and Abbott, R. D. (1984), “On errors-in-variables for binary regression models,” *Biometrika*, 71, 19–25.
- Castelnuovo, B., Kiragga, A., Schaefer, P., Kambugu, A., and Manabe, Y. (2009), “High rate of misclassification of treatment failure based on WHO immunological criteria in resource limited settings in Uganda,” *AIDS*, 23, 1295–1296.
- Deeks, S. G., Barbour, J. D., Grant, R. M., and Martin, J. N. (2002), “Duration and predictors of CD4 T-cell gains in patients who continue combination therapy despite detectable plasma viremia,” *AIDS*, 16, 201–207.
- Deeks, S. G., Barbour, J. D., Martin, J. N., Swanson, M. S., and Grant, R. M. (2000), “Sustained CD4+ T cell response after virologic failure of protease inhibitor-based regimens in patients with human immunodeficiency virus infection,” *Journal of Infectious Diseases*, 181, 946–953.

- Efron, B. (1981), “Nonparametric standard errors and confidence intervals,” *Canadian Journal of Statistics*, 9, 139–158.
- Fiscus, S., Cheng, B., Grove, S., Dementier, L., Jennings, C., Miller, V., and et al (2006), “HIV-1 viral load assays for resource limited settings.” *PLoS Medicine*, 3, e417.
- Foulkes, A. S., Azzoni, L., Li, X., Johnson, M., Smith, C., Mounzer, K., and Montaner, L. (2010), “Prediction based classification for longitudinal biomarkers,” *The Annals of Applied Statistics*, 4, 1476–1497.
- Foulkes, A. S. and De Gruttola, V. (2002), “Characterizing the Relationship between HIV-1 Genotype and Phenotype: Prediction-Based Classification,” *Biometrics*, 58, 145–156.
- Freund, Y., Schapire, R., and Abe, N. (1999), “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, 14, 1612.
- Fuller, W. (2009), *Measurement error models*, vol. 305, Wiley.
- Gardner, E., Burman, W., Steiner, J.F., A. P., and Bangsberg, D. (2009), “Antiretroviral medication adherence and the development of class-specific antiretroviral resistance,” *AIDS*, 23, 1035–1046.
- Gatsonis, C. (2009), “Receiver Operating Characteristic Analysis for the Evaluation of Diagnosis and Prediction1,” *Radiology*, 253, 593–596.
- Gillani, F. S. (2009), “The Miriam Hospital Immunology Center Database (ICDB) Annual Data Report,” Tech. rep., <http://www.rhodeislandhospital.org/cfar/ICDB-guidelines.htm>.
- Greengrass, V., Lohman, B., Morris, L., Plate, M., Steele, P. M., Walson, J. L., and Crowe, S. M. (2009), “Assessment of the low-cost Cavid ExaVir Load assay for monitoring HIV viral load in pediatric and adult patients.” *JAIDS*, 52, 387–390.
- Hagan, M., Demuth, H., Beale, M., et al. (1996), *Neural network design*, Thomson Learning Stamford, CT.
- Hammer, S., Saag, M., Schechter, M., Montaner, J., Schooley, R., Jacobsen, D., and et al (2006), “Treatment for adult HIV infection: 2006 recommendations of the international AIDS society-US panel.” *JAMA*, 296, 827–843.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Springer-Verlag Inc.
- Hsieh, F. and Turnbull, B. W. (1996), “Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve,” *The Annals of Statistics*, 24, 25–40.
- Huang, Y., Pepe, M., and Feng, Z. (2007), “Evaluating the predictiveness of a continuous marker.” *Biometrics*, 63, 1181–1188.

- Joska, J., Westgarth-Taylor, J., Hoare, J., Thomas, K., Paul, R., Myer, L., and Stein, D. (2011), "Validity of the International HIV Dementia Scale in South Africa." *AIDS Patient Care STDS*, 25(2), 95–101.
- Justice, A., McGinnis, K., Skanderson, M., Chang, C., Gibert, C., Goetz, M., Rimland, D., Rodriguez-Barradas, M., Oursler, K., Brown, S., et al. (2010), "Towards a combined prognostic index for survival in HIV infection: The role of 'non-HIV' biomarkers," *HIV Medicine*, 11, 143–151.
- Kantor, R., Diero, L., DeLong, A., Kamle, L., Muyonga, S., Mambo, F., Walumbe, E., Emonyi, W., Chan, P., Carter, E. J., Hogan, J., and Buziba, N. (2009), "Misclassification of First-Line Antiretroviral Treatment Failure Based on Immunological Monitoring of HIV Infection in Resource-Limited Settings," *Clinical Infectious Diseases*, 49, 454–462.
- Keiser, O., Macphail, P., Boule, A., Wood, R., Schechter, M., Dabis, F., and et al (2009), "Accuracy of WHO CD4 cell count criteria for virological failure of antiretroviral therapy," *Tropical Medicine and International Health*, 14, 1220–1225.
- Kenya (2005), *Guidelines to Antiretroviral Drug Therapy in Kenya. National AIDS and STD control Programme, 3rd Ed*, Kenya Ministry of Health.
- Kiragga, A. N., Castelnovo, B., Kamya, M. R., Moore, R., and Manabe, Y. C. (2012), "Regional differences in predictive accuracy of WHO immunologic failure criteria," *AIDS*, 26, 768–770.
- Kotsiantis, S. B. (2007), "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, 31, 249–268.
- Lee, A. (1990), *U-Statistics: Theory and Practice*, Marcel Dekker, New York.
- Lynen, L., An, S. and Koole, O., Thai, S., Ros, S., De Munter, P., and et al (2009), "An algorithm to optimize viral load testing in HIV-positive patients with suspected first-line antiretroviral therapy failure in Cambodia," *JAIDS*, 52, 40–48.
- Mabeya, H., Khozaim, K., Liu, T., Orango, O., Chumba, D., Pisharodi, L., Carter, J., and Cu-Uvin, S. (2012), "Comparison of Conventional Cervical Cytology Versus Visual Inspection With Acetic Acid Among Human Immunodeficiency Virus–Infected Women in Western Kenya," *Journal of Lower Genital Tract Disease*, 16, 92–97.
- Malawi (2003), *Treatment of AIDS, Guidelines for the use of Antiretroviral therapy in Malawi, 1st Ed*, Malawi Ministry of Health.
- Mee, P., Fielding, K., Charalambous, S., Churchyard, G., and AD, G. (2008), "Evaluation of the WHO criteria for antiretroviral treatment failure among adults in South Africa," *AIDS*, 22(15), 1971–1977.
- Metz, C. . E. (1978), "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, 8, 283–298.

- Meya, D., Spacek, L., Tibenderana, H., John, L., Namugga, I., Magero, S., and et al (2009), “Development and evaluation of a clinical algorithm to monitoring patients on antiretrovirals in resource-limited settings using adherence, clinical and CD4 cell count criteria,” *Journal of the International AIDS Society*, 12, 3.
- Moore, D. M., Hogg, R. S., Yip, B., Wood, E., Tyndall, M., Braitstein, P., and Montaner, J. S. (2005), “Discordant immunologic and virologic responses to highly active antiretroviral therapy are associated with increased mortality and poor adherence to therapy,” *JAIDS*, 40, 288–293.
- Pahwa, S., Read, J. S., Yin, W., Matthews, Y., Shearer, W., Diaz, C., Rich, K., Mendez, H., Thompson, B., for the Women, and Study, I. T. (2008), “CD4+/CD8+ T Cell Ratio for Diagnosis of HIV-1 Infection in Infants: Women and Infants Transmission Study,” *Pediatrics*, 122, 331–339.
- Panel on Antiretroviral Guidelines for Adults and Adolescents (2011), *Guidelines for the use and antiretroviral agents in HIV-1-infected adults and adolescents.*, Department of Health and Human Services (DHHS), WHO.
- Parmigiani, G. (2002), *Modeling in Medical Decision Making: A Bayesian Approach*, Wiley Series in Statistics in Practice.
- Pepe, M. (2011), “Problems with risk reclassification methods for evaluating prediction models,” *American Journal of Epidemiology*, 173, 1327–1335.
- Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R. L., Thompson, I., and et al (2008), “Inteprating the predictiveness of a marker with its performance as a classifier.” *American Journal of Epidemiology*, 167, 362–368.
- Pepe, M. S. (2000), “Receiver Operating Characteristic Methodology,” *Journal of the American Statistical Association*, 95, 308–311.
- (2003), *The statistical evaluation of medical tests for classification and prediction.*, Oxford University Press.
- Pepe, M. S. and Thompson, M. L. (2000), “Combining diagnostic test results to increase accuracy,” *Biostatistics*, 1, 123–140.
- Prentice, R. L. and Pyke, R. (1979), “Logistic disease incidence models and case-control studies.” *Biometrika*, 66, 403–11.
- Qin, J. (1999), “Empirical Likelihood Ratio Based Confidence Intervals for Mixture Proportions,” *The Annals of Statistics*, 27, 1368–1384.
- Republic Zambia (2004), *National Guidelines on Management and Care of Patients with HIV/AIDS*, Republic Zambia Ministry of Health.
- Reynolds, S., Nakigozi, G., Newell, K., Ndyababo, A., Galiwongo, R., Boaz, I., Quinn, T., Gray, R., Wawer, M., and Serwadda, D. (2009), “Failure of immunologic criteria to appropriately identify antiretroviral treatment failure in Uganda,” *AIDS*, 23(6), 697–700.

- Rosner, B., Spiegelman, D., and Willett, W. (1990), “Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error,” *American Journal of Epidemiology*, 132, 734–745.
- Sarle, W. (ed.) (1994), *Neural networks and statistical models*, In: Proceedings of the Nineteenth Annual SAS Users Group International Conference, pp. 1538–1550. SAS Institute.
- Schechter, M. and Tuboi, S. H. (2006), “Discordant immunological and virologic responses to antiretroviral therapy,” *Journal of Antimicrobial Chemotherapy*, 58, 506–510.
- Schooley, R. (2007), “Viral load testing in resource-limited settings,” *Clinical Infectious Diseases*, 44, 139–140.
- Sinisi, S. E., Polley, E. C., Petersen, M. L., Rhee, S.-Y., and van der Laan, M. J. (2007), “Super learning: an application to the prediction of HIV-1 drug resistance,” *Statistical Applications in Genetics and Molecular Biology*, 6, 7.
- Stefanski, L. and Cook, J. (1995), “Simulation-extrapolation: the measurement error jackknife,” *Journal of the American Statistical Association*, 90, 1247–1256.
- Steyerberg, E., Vickers, A., Cook, N., and et al (2010), “Assessing the performance of prediction models: a framework for traditional and novel measures,” *Epidemiology*, 21, 128–138.
- Thompson, M. A., Aberg, J. A., Cahn, P., Montaner, J. S. G., Rizzardini, G., Telenti, A., Gatell, J. M., Günthard, H. F., Hammer, S. M., Hirsch, M. S., Jacobsen, D. M., Reiss, P., Richman, D. D., Volberding, P. A., Yeni, P., and Schooley, R. T. (2010), “Antiretroviral Treatment of Adult HIV Infection,” *JAMA*, 304, 321–333.
- Tuboi, S. H., Brinkhof, M. W., Egger, M., Stone, R. A., Braitstein, P., Nash, D., Sprinz, E., Dabis, F., Harrison, L. H., and Schechter, M. (2007), “Discordant responses to potent antiretroviral therapy in previously naive HIV-1-infected adults initiating therapy in resource-constrained countries: The antiretroviral therapy in low-income countries (ART-LINC) collaboration,” *JAIDS*, 45, 52–59.
- Uganda (2003), *National Antiretroviral Treatment and Care Guidelines for Adults and Children (1st Edition)*, Uganda Ministry of Health.
- UNAIDS (2009), *2009 AIDS Epidemic Update* (http://data.unaids.org/pub/Report/2009/2009_epidemic_update_en.pdf), World Health Organization (WHO).
- (2010), *Report on the global AIDS epidemic*, World Health Organization (WHO).
- van der Laan, M. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer.
- van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.
- Vekemans, M., John, L., and Colebunders, R. (2007), “When to switch for antiretroviral treatment failure in resource-limited settings?” *AIDS*, 21, 1205–1206.

Vickers, A. and Elkin, E. (2006), “Decision curve analysis: A novel method for evaluating prediction models,” *Medical Decision Making*, 26, 565–574.

WHO (2010a), *Antiretroviral therapy for HIV infection in adults and adolescents: 2010 Revision.*, World Health Organization, Geneva.

— (2010b), *Towards universal access: Scaling up priority HIV/AIDS interventions in the health sector*, World Health Organization, Geneva.

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

APPENDIX A.

A.1 Semiparametric estimates of G_0 , G_1 , and G under the exponential tilt assumption

Suppose that we want to estimate the mixture distribution G using an i.i.d sample of $\{(S_i, Z_i) : i = 1, \dots, n\}$. In the spirit of nonparametric likelihood estimation, we consider only the distributions with jumps at $\{S_i\}$. Thus the (profile) likelihood for G_0 can be written as (see Qin 1999)

$$\begin{aligned} \mathbf{L}(G_0) &\propto \prod_{\{Z_i=0\}} g_0(S_i) \prod_{\{Z_i=1\}} \exp(\hat{\beta}_0^* + \hat{\beta}_1 S_i) g_0(S_i) \\ &= \left\{ \prod_{i=1}^n \theta_i \right\} \prod_{\{Z_i=1\}} \exp(\hat{\beta}_0^* + \hat{\beta}_1 S_i), \end{aligned}$$

where $\hat{\beta}_0^* = \hat{\beta}_0 - \text{logit}(\hat{p})$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the MLEs from the logistic regression (6), and $\theta_i = g_0(S_i)$ denotes the mass at the observed S_i with $\sum_i \theta_i = 1$. Here we proceed as if we have n distinct values in $\{S_i\}$, which does not affect the following results. Applying the Lagrange multiplier, one can show that the likelihood is maximized at

$$\hat{\theta}_i = \frac{1}{n} \left[1 + \nu \{ \exp(\hat{\beta}_0^* + \hat{\beta}_1 S_i) - 1 \} \right]^{-1},$$

where ν is the Lagrange multiplier solving

$$\sum_{i=1}^n \frac{\exp(\hat{\beta}_0^* + \hat{\beta}_1 S_i) - 1}{1 + \nu \{ \exp(\hat{\beta}_0^* + \hat{\beta}_1 S_i) - 1 \}} = 0.$$

We then estimate G_0 , G_1 , and G semiparametrically by

$$\begin{aligned} \tilde{G}_0(s) &= \sum_{i=1}^n \hat{\theta}_i \mathbf{1}\{S_i < s\}, \\ \tilde{G}_1(s) &= \sum_{i=1}^n e^{(\hat{\beta}_0^* + \hat{\beta}_1 S_i)} \hat{\theta}_i \mathbf{1}\{S_i < s\}, \\ \tilde{G}(s) &= (1 - \hat{p}) \tilde{G}_0 + \hat{p} \tilde{G}_1. \end{aligned}$$

Because the exponential tilt assumption places no constraint on the marginal distribution G , it can be verified that the semiparametric estimate $\tilde{G}(s)$ is equal to the empirical estimate $\hat{G}(s)$.

A.2 Properties of AUC_ϕ for tripartite rules

Property A.2.1: Let $S \sim G_1$, $S' \sim G_0$, and S and S' be independent. Then,

$$AUC_\phi = \Pr\{S > H_\phi(S')\} + \frac{1}{2} \Pr\{S = H_\phi(S')\}, \quad (\text{A.1})$$

and

$$AUC_\phi = \frac{1}{2} (1 + E_{G_1}[G_0(S)] - E_{G_0}[G_1\{H_\phi(S')\}]) . \quad (\text{A.2})$$

Proof: We have

$$\begin{aligned} AUC_\phi &= \int_{-\infty}^{\infty} \{1 - G_1(H_\phi(u))\} d(1 - G_0(u)) \\ &= \int_{-\infty}^{\infty} \Pr(S > H_\phi(u)) dG_0(u) \\ &\Rightarrow (\text{A.1}), \end{aligned}$$

where $\frac{1}{2} \Pr(X = H_\phi(X'))$ is added for ties and the term vanishes for continuous S . Further,

$$\begin{aligned} AUC_\phi &= 1 - \int_{-\infty}^{\infty} G_1(H_\phi(u)) dG_0(u) \\ &= \int_{-\infty}^{\infty} G_0(u) dG_1(H_\phi(u)) \\ &\Rightarrow (\text{A.2}). \end{aligned}$$

□

Property A.2.2: If S is stochastically greater than S' , then AUC_ϕ is bounded by

$$\frac{1}{2} + \phi - \frac{\phi^2}{2} \leq AUC_\phi \leq 1.$$

The lower bound is achieved when $G_1 = G_0$, and the upper bound when $\phi = 1$.

Proof: We prove the results for the case when S is continuous such that there exist l and u with $G(u) - G(l) = \phi$. Manipulating this constraint slightly, we have $1 - G(l) = 1 - G(u) + \phi \Rightarrow 1 - \{pG_1(H_\phi(u)) + (1 - p)G_0(H_\phi(u))\} = 1 - \{pG_1(u) + (1 - p)G_0(u)\} + \phi$. The condition that S is stochastically greater than S' implies that

$$\begin{aligned} 1 - G_1(H_\phi(u)) &\geq 1 - \{pG_1(u) + (1 - p)G_0(u)\} + \phi \\ &\geq 1 - G_0(u) + \phi. \end{aligned}$$

Therefore,

$$\begin{aligned} AUC_\phi &= \int_{u=-\infty}^{\infty} \{1 - G_1(H_\phi(u))\} d(1 - G_0(u)) \\ &\geq \int_{u=-\infty}^{\infty} [\{1 - G_0(u) + \phi\} \wedge 1] d(1 - G_0(u)) \\ &= \int_{u=-\infty}^{G_0^{-1}(\phi)} \{1 - G_0(u) + \phi\} d(1 - G_0(u)) + \int_{u=G_0^{-1}(\phi)}^{\infty} d(1 - G_0(u)) \\ &= \frac{1}{2} + \phi - \frac{\phi^2}{2}. \end{aligned}$$

All equalities hold when $G_0 = G_1$. \square

A.3 Asymptotic properties of estimated ROC curve and AUC

The nonparametric estimate \widehat{C}_ϕ given by (9) has the following properties:

Property A.3.1: *The nonparametric estimate \widehat{C}_ϕ is uniformly consistent.*

Proof: Let us write

$$\begin{aligned} \sup_t |\widehat{C}_\phi(t) - C_\phi(t)| &= \sup_t |\widehat{G}_1 \circ \widehat{H}_\phi \circ \widehat{G}_0^{-1}(t) - G_1 \circ \widehat{H}_\phi \circ \widehat{G}_0^{-1}(t)| \\ &\quad + \sup_t |G_1 \circ \widehat{H}_\phi \circ \widehat{G}_0^{-1}(t) - G_1 \circ H_\phi \circ \widehat{G}_0^{-1}(t)| \\ &\quad + \sup_t |G_1 \circ H_\phi \circ \widehat{G}_0^{-1}(t) - G_1 \circ H_\phi \circ G_0^{-1}(t)|. \end{aligned}$$

Then, it can be shown that the first term converges to zero almost surely by the Glivenko-Cantelli Theorem, and the second and third terms converge to zero almost surely by the Law of Large Numbers. See Hsieh and Turnbull (1996). \square

Property A.3.2: *Suppose that the densities g_0 , g_1 and g are continuous and bounded, and $\sum Z_i/n \rightarrow p$ as $n \rightarrow \infty$. Then, the following approximation holds asymptotically as $n \rightarrow \infty$,*

$$\begin{aligned} n^{\frac{1}{2}}\{\widehat{C}_\phi(v) - C_\phi(v)\} &= \frac{1}{\sqrt{p}}B_1 \circ C_\phi(v) + \frac{g_1 \circ H \circ G_0^{-1}(1-v)\sqrt{1-p}}{g \circ G^{-1}\{G \circ G_0^{-1}(1-v) - \phi\}}B\{G \circ G_0^{-1}(1-v) - \phi\} \\ &\quad + \frac{1}{\sqrt{1-p}}\frac{g_1 \circ H \circ G_0^{-1}(1-v)}{g_0 \circ G_0^{-1}(1-v)}\frac{g \circ G_0^{-1}(1-v)}{g \circ G^{-1}\{G \circ G_0^{-1}(1-v) - \phi\}}B_2(1-v), \end{aligned}$$

where $B_1(v)$ and $B_2(v)$ are independent Brownian bridges, $B(v) = B_1(v)/\sqrt{p} + B_2(v)/\sqrt{1-p}$, and $\rho = \{G \circ G_0^{-1}(v) - \phi\}\{1 - G \circ G_0^{-1}(v)\}/[G \circ G_0^{-1}(v)\{1 - G \circ G_0^{-1}(v) + \phi\}]$.

The strategy for proving A.3.2 is similar to Hsieh and Turnbull (1996).

Property A.3.3: *The nonparametric estimate $\widehat{\text{AUC}}_\phi$ given by (10) has the property that, as $n \rightarrow \infty$,*

$$n^{1/2}\sigma^{-1/2}(\widehat{\text{AUC}}_\phi - \text{AUC}_\phi) \xrightarrow{d} \mathcal{N}(0, 1), \quad (\text{A.3})$$

where $\sigma^2 = \text{Var}_{G_1}\{G_0 \circ H_\phi(S)\}p^{-1} + \text{Var}_{G_0}\{G_1 \circ H_\phi(S)\}(1-p)^{-1}$.

Proof: We prove (A.3) using the properties of U-statistics (Lee 1990). Applying the Hájek projection principle on (10) (van der Vaart 1998), we express $\widehat{\text{AUC}}_\phi$ as

$$\widehat{\text{AUC}}_\phi = \text{AUC}_\phi + \widetilde{A}_n + o(1/n),$$

where

$$\begin{aligned} \widetilde{A}_n &= \frac{1}{\sum_i Z_i} \sum_i Z_i [\widehat{G}_0 \circ \widehat{H}_\phi(S_i) - \text{E}_{\widehat{G}_1}\{\widehat{G}_0 \circ \widehat{H}_\phi(S_i)\}] \\ &\quad + \frac{1}{\sum_i (1 - Z_i)} \sum_i (1 - Z_i) [\widehat{G}_1 \circ \widehat{H}_\phi(S_i) - \text{E}_{\widehat{G}_0}\{\widehat{G}_1 \circ \widehat{H}_\phi(S_i)\}] \end{aligned}$$

is a U-statistic. Then conditional on $\sum_i Z_i$, an ancillary statistic for the AUC, (A.3) is an immediate result of applying Slutsky's lemma and the Central Limit theorem. \square