Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 97

Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections

Tianxi Cai* Lu Tian†
Peggy H. Wong‡ L. J. Wei**

*Harvard University, tcai@hsph.harvard.edu
†Stanford University School of Medicine, lutian@stanford.edu
‡Merck Research Laboratories, peggy_wong@merck.com
**Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper97

ANALYSIS OF RANDOMIZED COMPARATIVE CLINICAL TRIAL DATA FOR PERSONALIZED TREATMENT SELECTIONS

By Tianxi Cai

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.

tcai@hsph.harvard.edu

Lu Tian

Department of Health Policy and Research, Stanford University School of Medicine, Stanford, California 94305, U.S.A.

lutian@stanford.edu

Peggy H. Wong

Merck Research Laboratories , Rahway NJ 07065, U.S.A.

peggy_wong@merck.com

and L.J. Wei

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.

wei@hsph.harvard.edu

SUMMARY

Suppose that under the conventional randomized clinical trial setting, a new therapy is compared with a standard treatment. In this article, we propose a systematic, two-stage estimation procedure for the subject-level treatment differences for future patient's disease management and treatment selections. To construct this procedure, we first utilize a parametric or semi-parametric method to estimate individual-level treatment differences and use these estimates to create an index scoring system for clustering patients. We then consistently estimate the average treatment difference for each cluster of subjects via a nonparametric function estimation method. Furthermore, pointwise and simultaneous interval estimates are constructed to make inferences about such individual-specific treatment differences. The new proposal is illustrated with the data from a clinical trial for evaluating the efficacy and toxicity of a three drug combination vs. a standard two drug combination for treating HIV-1 infected patients.

Keywords: Cross-validation; HIV-infection; Nonparametric function estimation; Personalized medicine; Subgroup analysis.

1. INTRODUCTION

One of the major components for modern evidence-based medicine is to utilize the patient's "baseline" information for personalized disease management and treatment selection. For instance, in a recent study it was demonstrated that the benefit of giving toxic chemotherapy prior to hormone therapy with tamoxifen for postmenopausal women with lymph node-negative breast cancer varies depending on the estrogen receptor (ER) status of the tumor. Those with ER-negative tumors benefited substantially from chemotherapy whilst those with ER-positive tumors did not benefit as compared to receiving tamoxifen alone (IBCSG, 2002). In another example with an observational study, Sabine (2005) suggested that the efficacy and toxicity profiles for the highly active antiretroviral therapy (HAART) vary markedly across different subgroups of HIV infected patients and recommended certain subject-specific treatment strategies. These individualized decision rules can be extremely useful in practice. However, there are numerous examples that the results from the so-called subgroup analyses, which were not properly planned or executed subgroup analyses, could not be validated (Rothwell, 2005; Pfeffer & Jarcho, 2006; Wang et al., 2007).

In this paper, we consider the case that a new therapy was compared with a control under the standard randomized comparative clinical trial setting. Generally the main goal of a randomized clinical trial is to make inferences about an *overall* treatment difference with respect to efficacy and toxicity. On the other hand, a "positive" trial does not imply that all future patients would benefit from the new treatment. Moreover, a "negative" study does not mean all patients should be treated by the standard therapy. In fact, based on the extensive collection of the study patient's baseline information from a clinical trial, it would be valuable to utilize such information to make inferences about the individual-level treatment efficacy.

In practice, a commonly employed first step to perform subgroup analyses is to examine whether there are statistically significant interactions between the treatment assignment indicator and various baseline covariates via, for example, parametric regression models (Byar, 1985). This ad hoc "fishing expedition" approach may not accurately or efficiently identify proper subgroups of patients who would benefit from the new treatment. Recently, Song & Pepe (2004) proposed a novel procedure to identify a critical value for a *single* covariate, which may guide us to split the future population into two groups. Patients in one group would be treated by the new therapy and those

in the other group would take the control. Although such a approach is interesting with respect to an overall utility for the entire population, it does not provide a treatment choice scheme at a subject-specific level. Moreover, if the treatment difference is not monotonically associated with this single covariate, such a procedure may not be appropriate. To examine the treatment difference at a subject-level with a *single* covariate, Bonetti & Gelber (2000; 2005) clustered the patients with their covariate values and utilized a moving average procedure to make inferences about the profile of the treatment differences over the covariate. Recently, in her unpublished thesis, Park (2004) proposed a procedure to identify patients who may or may not benefit from the new treatment based on generalized linear models and the Cox proportional hazards models. The validity of her procedure heavily relies on the model assumptions.

In this paper, we consider the case that each patient has *multiple* baseline covariates and propose a systematic, two-stage method to identify patients who would benefit from the new treatment. Specifically, for the first stage, we use a parametric or semi-parametric model to estimate the subject-specific mean response for each treatment group. We then utilize the resulting difference of these two estimates as an index scoring system to cluster patients. That is, subjects in each cluster would have the same parametric index score. For the second stage, we calibrate the estimated treatment differences with a consistent, nonparametric function estimation procedure and provide valid *pointwise* and *simultaneous* inferences about the true average treatment difference for each cluster of patients by controlling the desirable confidence level locally and globally . The new proposal is illustrated with a data set from a clinical trial for evaluating a three-drug combination vs. a standard two-drug combination for treating HIV-infected patients. For cost-benefit decision makings, we also provide the underlying mean treatment response for each treatment group over the index score and the estimated relative frequency for the parametric score.

2. POINT AND INTERVAL ESTIMATION FOR TREATMENT DIFFERENCES

Suppose that study subjects in a population of interest are randomly assigned to two treatment groups, $\{G_k, k = 0, 1\}$. Let Y_k and $\mathbf{U}_k$ denote the response Y and the covariate vector $\mathbf{U}$ for the kth group, respectively. Now, suppose that for subjects with $\mathbf{U} = \mathbf{u}$, we are interested in estimating the treatment difference

$$\mathcal{S}(\mathbf{u}) = E(Y_1 - Y_0 \mid \mathbf{U}_1 = \mathbf{U}_0 = \mathbf{u}).$$

Our data consist of $\{(Y_{ki}, \mathbf{U}_{ki}), i = 1, ..., n_k\}$, n_k independent and identical copies of $(Y_k, \mathbf{U}_k), k = 0, 1$. Assume that as $n_0 \to \infty$, the ratio n_1/n_0 goes to a constant in the open interval $(0, 1)$. To estimate $\mathcal{S}(\mathbf{u})$, one may consider a non-parametric function procedure. In practice, however, when $\mathbf{u}$ is not univariate, generally it is difficult, if not impossible, to estimate $\mathcal{S}(\mathbf{U})$ well non-parametrically.

To reduce the complexity of the high dimensional problem, conventionally one utilizes a parametric or semi-parametric procedure to estimate $\mathcal{S}(\mathbf{u})$. For the present case, we consider the following generalized linear *working* model to approximate the mean of Y_k with a function of $\mathbf{U}_k$:

$$E(Y_k \mid \mathbf{U}_k) = g_k(\boldsymbol{\beta}_k^{\mathsf{T}} \mathbf{Z}_k), \quad k = 0, 1, \tag{2·1}$$

where $\mathbf{Z}_k$, a $p \times 1$ vector, is a function of $\mathbf{U}_k$ with first column being 1, g_k is a known, strictly increasing link function, and $\boldsymbol{\beta}_k$ is an unknown vector of regression coefficients. Note that for (2.1), we only model the mean function of Y. To estimate the parameter vector $\boldsymbol{\beta}_k$ in (2.1) without distribution assumptions about the response, one may use a solution $\widehat{\boldsymbol{\beta}}_k$ to the following estimating equation

$$\sum_{i=1}^{n_k} \mathbf{Z}_{ki} \{Y_{ki} - g_k(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{ki})\} = 0. \tag{2·2}$$

Using the arguments given in Tian et al. (2007), one can show that $\widehat{\boldsymbol{\beta}}_k$ converges to a deterministic vector $\bar{\boldsymbol{\beta}}_k$ even when Model (2.1) is incorrectly specified. This stability property is crucial for developing our new procedure. It follows that for a given $\mathbf{u}$ or $\mathbf{z}$, a parametric estimator for $\mathcal{S}(\mathbf{u})$ is

$$\widehat{s}(\mathbf{u}) = g_1(\widehat{\boldsymbol{\beta}}_1^{\mathsf{T}} \mathbf{z}) - g_0(\widehat{\boldsymbol{\beta}}_0^{\mathsf{T}} \mathbf{z}).$$

Note that when Model (2.1) is correctly specified, $\bar{\boldsymbol{\beta}}_k$ is the true parameter for Model (2.1) and $\widehat{s}(\mathbf{u})$ is a consistent estimator of $\mathcal{S}(\mathbf{u})$.

Let $\mathbf{U}^0$ be a typical baseline covariate vector for a future subject from the study population. If this subject is treated by treatment k, the response is $Y_k^0, k = 0, 1$. For $\mathbf{U}^0 = \mathbf{u}^0$, we may use $\widehat{s}(\mathbf{u}^0)$ to decide which treatment this specific subject should be treated with. The adequacy of such a decision heavily depends on the appropriateness of Model (2.1). On the other hand, $\widehat{s}(\cdot)$ may be used as an index scoring system for clustering future subjects with potentially similar treatment differences. That is, we divide the future population into many strata based on the score $\widehat{s}(\cdot)$ such that patients in the same stratum have the same parametric score value.

Now, consider subjects in a stratum such that $\widehat{s}(\mathbf{U}^0) = v$, a given value, we are interested in consistently estimating the average treatment difference

$$\bar{\Delta}(v) = \mu_1(v) - \mu_2(v),$$

where $\mu_k(v) = E(Y_k^0 \mid \widehat{s}(\mathbf{U}^0) = v), k = 0, 1$, and the expectation is taken with respect the data and $(Y^0, \mathbf{U}^0)$. To estimate $\bar{\Delta}(v)$, we utilize a nonparametric function estimation procedure for $\mu_k(v)$ with a local likelihood score function (Tibshirani & Hastie, 1984; Fan & Gijbels, 1996). Specifically, we obtain the root $\{\widehat{a}_k(v), \widehat{b}_k(v)\}$ to the local weighted estimating equation, $\widehat{\mathbf{S}}_{kv}(a, b) = 0$, where

$$\widehat{\mathbf{S}}_{kv}(a, b) = \sum_{i=1}^{n_k} \begin{pmatrix} 1 \\ h^{-1}\widehat{\mathcal{E}}_{kvi} \end{pmatrix} K_h(\widehat{\mathcal{E}}_{kvi}) \left\{ Y_{ki} - \mathfrak{g}(a + b\widehat{\mathcal{E}}_{kvi}) \right\}, \tag{2·3}$$

h is the smoothing parameter, $K_h(x) = K(x/h)/h$, $K(x)$ is a symmetric kernel function with a finite support, $\widehat{\mathcal{E}}_{kvi} = \psi\{\widehat{s}(\mathbf{U}_{ki})\} - \psi(v)$ and $\psi(\cdot)$ is a known, non-decreasing function. Here, $\mathfrak{g}(x) = x$ if the response Y is continuous and $\mathfrak{g}(x) = \exp(x)/\{1 + \exp(x)\}$ if Y is binary. Note that we choose a transformation $\psi\{\widehat{s}(\mathbf{U}_{ki})\}$ of $\widehat{s}(\mathbf{U}_{ki})$ to implement the smoothing. In practice, a proper choice of $\psi(\cdot)$ can be critical (Wand et al., 1991; Park et al., 1997). The $\mu_k(v)$ can then be estimated by $\widehat{\mu}_k(v) = \mathfrak{g}\{\widehat{a}_k(v)\}$. This estimator corresponds to the local linear least square estimator for continuous Y and local linear logistic likelihood estimator for binary Y. Subsequently, we estimate $\bar{\Delta}(v)$ as

$$\widehat{\Delta}(v) = \widehat{\mu}_1(v) - \widehat{\mu}_0(v).$$

In Appendix A, for $h = O_p(n^{-\nu})$ with $1/5 < \nu < 1/2$, we show that $\widehat{\Delta}(v)$ is uniformly consistent for $\bar{\Delta}(v)$, for v in an interval which is properly contained in the support of $\widehat{s}(\mathbf{u})$.

For the above fixed value v, we show in Appendix B that with $h = O_p(n^{-\nu})$ and $1/5 < \nu < 1/2$, $\widehat{\mathcal{W}}(v) = (nh)^{\frac{1}{2}}\{\widehat{\Delta}(v) - \bar{\Delta}(v)\}$ is approximately normally distributed, where $n = n_0 + n_1$. Furthermore, we show that for large n, the distribution of $\widehat{\mathcal{W}}(v)$ can be approximated by the conditional distribution of a mean-zero normal variable

$$\begin{aligned} \widehat{\mathcal{W}}^*(v) =& (nh)^{\frac{1}{2}} \left[\frac{\sum_{i=1}^{n_1} K_h(\widehat{\mathcal{E}}_{1vi})\{Y_{1i} - \widehat{\mu}_1(v)\}\mathcal{Z}_{1i}}{\sum_{i=1}^{n_1} K_h(\widehat{\mathcal{E}}_{1vi})} - \frac{\sum_{i=1}^{n_0} K_h(\widehat{\mathcal{E}}_{0vj})\{Y_{0j} - \widehat{\mu}_0(v)\}\mathcal{Z}_{0j}}{\sum_{j=1}^{n_0} K_h(\widehat{\mathcal{E}}_{0vj})} \right] \\ &+ (nh)^{\frac{1}{2}} \left\{ \widehat{\Delta}(v, \widehat{\boldsymbol{\beta}}_1^*, \widehat{\boldsymbol{\beta}}_0^*) - \widehat{\Delta}(v) \right\}, \end{aligned} \tag{2·4}$$

given the data, where $\underline{\mathcal{Z}} = \{\mathcal{Z}_{ki}, i = 1, \cdots, n_k, k = 0, 1\}$ is a random sample from the standard normal variable and is independent of the data, $\widehat{\Delta}(v, \widehat{\boldsymbol{\beta}}_1^*, \widehat{\boldsymbol{\beta}}_0^*)$ is obtained by replacing $\widehat{\boldsymbol{\beta}}_k$ in $\widehat{\Delta}(v)$ with $\widehat{\boldsymbol{\beta}}_k^* = \widehat{\boldsymbol{\beta}}_k + \{\sum_{i=1}^{n_k} \dot{g}_k(\widehat{\boldsymbol{\beta}}_k^{\mathsf{T}}\mathbf{Z}_{ki})\mathbf{Z}_{ki}\mathbf{Z}_{ki}^{\mathsf{T}}\}^{-1}[\sum_{i=1}^{n_k} \mathbf{Z}_{ki}\{Y_{ki} - g_k(\widehat{\boldsymbol{\beta}}_k^{\mathsf{T}}\mathbf{Z}_{ki})\}\mathcal{Z}_{ki}]$, for $k = 0, 1$, where $\dot{g}(\cdot)$ is the derivative of $g(\cdot)$. Note that $\widehat{\boldsymbol{\beta}}_k^*$ is a solution to the counterpart of estimating equation (2.2), which is perturbed with the same set of $\underline{\mathcal{Z}}$ in (2.4). Also, note that the $\underline{\mathcal{Z}}$ are the only random quantities in (2·4), whose distribution can be approximated easily by simulating $\underline{\mathcal{Z}}$ repeatedly. A $(1 - \alpha)$ pointwise confidence interval estimates for $\bar{\Delta}(v)$ can be constructed via this large sample approximation, which is $\widehat{\Delta} \pm d(nh)^{-\frac{1}{2}}\widehat{\sigma}(v)$. Here, $\widehat{\sigma}(v)$ is the standard error estimate of $\widehat{\mathcal{W}}^*(v)$ and d is the upper $(1 - \alpha/2)$ percentile of the standard normal.

To control the global error rate, one may construct a simultaneous confidence band. To make inference about the treatment differences over a range of v, one may construct a simultaneous confidence band for $\{\bar{\Delta}(v), v \in \mathcal{J} = [\rho_l, \rho_r]\}$, which is properly contained in the support of $\widehat{s}(\cdot)$. A conventional way to obtain such a confidence band is based on a sup-type statistic

$$\widehat{\mathcal{M}} = \sup_{v \in \mathcal{J}} \left| \widehat{\sigma}(v)^{-1} \widehat{\mathcal{W}}(v) \right|. \tag{2·5}$$

However, as a process in v, $\widehat{\mathcal{W}}(v)$ does not converge weakly to a proper stochastic process, as $n \to \infty$. Therefore, we cannot use the standard large sample theory for empirical processes to obtain a finite sample approximation to the distribution of $\widehat{\mathcal{M}}$. On the other hand, by the strong approximation arguments and extreme value limit theorem (Bickel & Rosenblatt, 1973), in Appendix C, we show that a standardized version of $\widehat{\mathcal{M}}$ converges in distribution to a proper random variable. In practice, for large n, one can approximate the distribution of $\widehat{\mathcal{M}}$ by $\widehat{\mathcal{M}}^*$, the sup of the absolute value of $\widehat{\mathcal{W}}^*(v)$ divided by $\widehat{\sigma}(v)$, with (2·4) perturbed by the same set of perturbation variables $\underline{\mathcal{Z}}$ for all $v \in \mathcal{J}$. It follows that the $100(1 - \alpha)\%$ simultaneous confidence interval for $\bar{\Delta}(v)$ is

$$\widehat{\Delta}(v) \pm (nh)^{-1/2}\gamma\, \widehat{\sigma}(v), \tag{2·6}$$

where the cutoff point γ is chosen such that $\text{pr}(\widehat{\mathcal{M}}^* < \gamma) \geq 1 - \alpha$.

As for any nonparametric functional estimation problem, the choice of the smoothing parameter h for $\widehat{\Delta}(v)$ is crucial for making inferences about $\bar{\Delta}(v)$. It is important to note that for the present case, we need to choose a single smooth parameter for two nonparametric function estimates, $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$. Since each study subject is assigned to a single treatment group, therefore, we cannot use

the standard cross-validation method with the integrated mean square error criterion to choose h for $\widehat{\Delta}(v)$. A reasonable, feasible alternative to choose an "optimal" smooth parameter is minimizing an average squared distance between the observed cumulative treatment difference and their predicted counterparts in the validation samples under the $\mathcal{K}$-fold cross validation setting. Specifically, we randomly split the data into $\mathcal{K}$ disjoint subsets of about equal sizes, denoted by $\{\mathcal{J}_\iota, \iota = 1, \cdots, \mathcal{K}\}$, where $\mathcal{J}_\iota = \{\mathcal{J}_{\iota 1}, \mathcal{J}_{\iota 0}\}$ and $\mathcal{J}_{\iota k}$ denotes the collection of the observations in the ιth subset that belong to the kth treatment group, for $k = 0, 1$. For each ι, we use all observations not in $\mathcal{J}_\iota$ to obtain estimators for the parametric index score and $\hat{\Delta}(v)$ with a given h. Let the resulting estimators be denoted by $\widehat{s}_{(-\iota)}(\mathbf{U})$ and $\widehat{\Delta}_{(-\iota)}(v)$, respectively. We then use the observations from $\mathcal{J}_\iota$ to calculate the empirical integrated cumulative square error

$$\int \left\{ \frac{\sum_{i\in\mathcal{J}_{\iota 1}} I(\mathbf{U}_{1i} \le \mathbf{u}) Y_{1i}}{\sum_{i\in\mathcal{J}_{\iota 1}} I(\mathbf{U}_{1i} \le \mathbf{u})} - \frac{\sum_{j\in\mathcal{J}_{\iota 0}} I(\mathbf{U}_{0j} \le \mathbf{u}) Y_{0j}}{\sum_{j\in\mathcal{J}_{\iota 0}} I(\mathbf{U}_{0j} \le \mathbf{u})} - \frac{\sum_{k=0}^{1}\sum_{i\in\mathcal{J}_{\iota k}} I(\mathbf{U}_{ki} \le \mathbf{u}) \widehat{\Delta}_{(-\iota)}(\widehat{s}_{(-\iota)}(\mathbf{U}_{ki}))}{\sum_{k=0}^{1}\sum_{i\in\mathcal{J}_{\iota k}} I(\mathbf{U}_{ki} \le \mathbf{u})} \right\}^2 d\widehat{\mathcal{H}}(\mathbf{u}), \quad (2{\cdot}7)$$

where $I(\cdot)$ is the indicator function and $\widehat{\mathcal{H}}(\mathbf{u})$ is an empirical weight function such as the empirical distribution function. Lastly we sum (2·7) over $\iota = 1, \cdots, \mathcal{K}$, and then choose h by minimizing this sum. It is important to note that since the bandwidth is selected by minimizing a quantity which is composed of cumulative predicted errors, the order of such optimal bandwidth is expected to be $n^{-1/3}$ (Bowman et al., 1998). Thus the selected bandwidth satisfies the condition required for the resulting functional estimator $\widehat{\Delta}(v)$ with the data-dependent smooth parameter to have the above desirable large sample properties.

3. ESTIMATING TREATMENT DIFFERENCES FOR HIV-INFECTED PATIENTS

We use a data set from a clinical trial conducted by the AIDS Clinical Trials Group (ACTG) to illustrate the new proposal. Here, we present two cases, the first one is for a continuous response and the second case is for a binary endpoint. This study, ACTG 320, is one of early studies for examining the clinical added value from a protease inhibitor with two nucleoside analogues for treating human immunodeficiency virus type 1 infection (Hammer et al., 1997). A total of 1156 patients were randomized to two treatment groups. Here, treatment 0 is the two drug combination, zidovudine

and lamivudine, and treatment 1 is the three drug combination consisting of the above two and indinarvir. The study was terminated at the second formal interim analysis due to substantial *overall* improvements from the three drug combination over the standard two drug combination with respect to various study endpoints. However, even with this potent new treatment, some patients may not respond to the therapy, but instead suffer from non-trivial toxicity. Therefore, for future patients' disease management, it is important to predict patient's treatment responses based on certain "baseline" markers.

For the first example of the illustration, we let the response Y be the patient's change of CD4 count at Week 24 from the baseline level, an important endpoint for evaluating HIV treatments. In our analysis, we only considered subjects who had baseline covariate and week 24 CD4 count values ($n_0 = 429$, $n_1 = 427$). Here, the vector $\mathbf{u}$ of baseline covariates consists of CD4, $\log_{10}$RNA and age. To accommodate the potential non-linear relationship between CD4 and the response, we let $\mathbf{z}$ in Model (2.1) be the vector $(1, \mathrm{CD4}, \log(\mathrm{CD4}), \log_{10}\mathrm{RNA}, \mathrm{Age})^{\intercal}$. Furthermore, we let $g_0(\cdot)$ and $g_1(\cdot)$ be the identity function. The estimated regression coefficients for fitted models via the estimating equations (2·2) are summarized in Table 1(a). It appears that the two resulting fitted regression functions $\widehat{s}(\mathbf{u})$ are rather different, indicating that there are potential interactions between the baseline covariates and the treatment response. It is interesting to note that for the two drug combination group, only age is marginally significant. On the other hand, the baseline CD4 and RNA are highly associated with the response for the three drug combination group. Moreover, note that younger patients with higher baseline RNA tend to have high scores. For example, among the 44 patients with age below 40, baseline $\log_{10}$RNA more than 5 and baseline CD4 between 100 and 150, the average score is 95. On the other hand, among the 46 patients with age above 40, baseline CD4 above 150 and $\log_{10}$RNA below 5, the average score is only 25.

For the second step, we used this continuous score index $\widehat{s}(\mathbf{u})$ to group future patients and predict the true treatment difference $\bar{\Delta}(v)$ for patients with $\widehat{s}(\mathbf{U}^0) = v$ with $\widehat{\Delta}(v)$. Here, since the estimated score has a left skewed distribution, we let $\psi(v) = \log(-\log[\Phi\{(v-70)/25\}])$ be the transformation for choosing the smooth parameter h. The kernel function $K(\cdot)$ is the Epanechnikov kernel and the smoothing parameter h was obtained through a 10-fold cross validation as a minimizer for the criterion (2·7). Furthermore, in (2.7), the weight function $\widehat{\mathcal{H}}(\mathbf{u}) = \widehat{\mathcal{F}}(\mathbf{u}) I(\widehat{\mathcal{F}}(\mathbf{u}) \in [0.05, 0.95])$, where $\widehat{\mathcal{F}}(\mathbf{u}) = n^{-1} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I(\mathbf{U}_{ki} \leq \mathbf{u})$. The resulting bandwidth h 0.93 for the transformed score which

has a range of . Here, the confidence intervals for $\bar{\Delta}(v)$ were constructed for the transformed score $\psi(v) \in [-1.73, 1.12]$. This corresponds to $v \in [30, 94]$, which is $[\psi^{-1}\{\psi(\widehat{q}_{0.01})+h\}, \psi^{-1}\{\psi(\widehat{q}_{0.99})-h\}]$, where $\widehat{q}_p$ is the pth percentile of the observed $\widehat{s}(\mathbf{U})$. In Figure 1(c), we present the estimate $\widehat{\Delta}(v)$, the solid curve. This curve, in some region, seems markedly different from the 45-degree line. For example, subjects with score 60 have an estimated average treatment difference of 45, indicating that the parametric risk estimates need to be calibrated.

To make further inferences about $\bar{\Delta}(v)$, we approximated the distribution of the estimator $\widehat{\Delta}(v)$ via the aforementioned perturbation-resampling method (2·4) with 500 independent realized Normal samples of $\underline{\mathcal{Z}}$. In Figure 1(c), we present the resulting 0.95 pointwise intervals (bounded by the dotted curves) and its simultaneous band (gray area). For risk-cost-benefit decision makings, we also present the estimates $\hat{\mu}_k, k = 0, 1$, the underlying mean changes at Week 24 for both treatment groups in Figure 1(b). We present the relative frequency of patients in the study population based on the index score $\widehat{s}(\mathbf{u})$ in Figure 1(a). Since most patients have estimated scores between 50 and 90, the estimation of the true treatment differences tends to be more precise in this region.

The average treatment response from the three drug combination group is uniformly higher than its counterpart from the two drug combination. However, the treatment differences do not change much when index score values are low, but for score values higher than 50, the treatment differences appear to increase significantly. The confidence interval estimates displayed in Figure 1(c) play important roles for treatment selections, especially with additional information regarding the toxicity profiles over the index score.

For the second example, we considered the patient's response being a binary variable, which is one if the week 24 RNA is below 500 copies per milliliter (the criterion for the RNA response used for ACTG 320). Since the linear effects of the baseline CD4 are almost 0 for both groups, we only included log(CD4) for this analysis. It follows that $\mathbf{z} = (1, \log(\text{CD4}), \log_{10}\text{RNA}, \text{Age})'$ and $g_0(x) = g_1(x) = \exp(x)/\{1 + \exp(x)\}$ for Model (2.1). Here, $n_0 = 408$ and $n_1 = 416$. In Table 1(b), we present the estimates for the regression parameters and their estimated standard errors. For the present case, the estimated score is between -1 and 1, we applied a transformation $\psi(x) = \log\{(x+1)/(1-x)\}$ for choosing smooth parameter. The optimal bandwidth corresponding to the transformed score is 0.49 using the 10-fold cross-validation procedure as the one for the continuous response case. The confidence intervals for $\bar{\Delta}(v)$ were constructed over $\psi(v) \in [0.72, 1.76]$

which corresponds to $v \in [0.34, 0.71]$. In Figure 2(c), we present the estimated treatment differences (solid curve) over the index score. For this endpoint, the three drug combination is substantially and uniformly better than the two drug combination as shown in Figure 2(b). For instance, for patients with a score of 0.35, the probability of having RNA suppressed below the limit of quantification is 46% if treated with the three-drug therapy and only 8% if treated with the two drug alternatives. For this subgroup, the true treatment difference was estimated as 0.38 with 95% pointwise confidence interval $[0.32, 0.43]$ and 95% simultaneous confidence interval $[0.30, 0.46]$.

4. REMARKS

Firstly, it is important to note that with the parametric estimate $\widehat{s}(\mathbf{u})$, the corresponding simultaneous confidence interval estimates for $\mathbf{S}(\mathbf{u})$, if valid, can be quite conservative since they would be obtained via a sup-statistic over $\mathbf{u}$, whose dimension can be rather large. Secondly, one may use the same approach presented in this article to estimate individual-specific treatment differences from an observational study. Such estimates, however, may not have the causal interpretation for the treatment intervention. Thirdly, the treatment difference may be quantified using other measures for the contrast between two treatment groups, for example, the relative risk or odds ratio for binary response variable. Fourthly, using the same idea presented in this article, one may develop subject-level inference procedures for the treatment differences with a censored event time endpoint. Lastly, an interesting and important question is how to evaluate the parametric index scoring system, which can "efficiently" cluster patients for the first stage of our proposal. Unlike the standard risk prediction problem, there is no obvious metric such as the receiver operating characteristic (ROC) curve to evaluate the performance of the index system globally or locally. On the other hand, heuristically one would choose a scoring system such that its frequency distribution of the resulting calibrated estimates for the treatment differences has wide spread over a large support.

Appendix

Throughout, unless noted otherwise, we use the notation $\simeq$ to denote equivalence up to $o_p(1)$ uniformly in v, $\lesssim$ to denote being bounded above up to a universal constant, and $\dot{\mathcal{F}}(x)$ to denote $d\mathcal{F}(x)/dx$ for any function $\mathcal{F}$. We use $\widehat{\mathbb{P}}_k$ and $\mathbb{P}_k$ to denote expectation with respect to the empirical probability measure of $\{(Y_{ki}, \mathbf{U}_{ki}), i = 1, \cdots, n_k\}$ and the probability measure of $(Y_k, \mathbf{U}_k)$, respectively. Similarly $\widehat{\mathbb{G}}_k = n^{\frac{1}{2}}(\widehat{\mathbb{P}}_k - \mathbb{P}_k)$.

Let $p_k = \lim_{n_0 \to \infty} n_k/n$, $\bar{\boldsymbol{\beta}}_k$ denote the solution to the equation $E[\mathbf{Z}_{ki}\{Y_{ki} - g_k(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}] = 0$, $\bar{s}(\mathbf{U}) = g_1(\bar{\boldsymbol{\beta}}_1'\mathbf{Z}) - g_0(\bar{\boldsymbol{\beta}}_0'\mathbf{Z})$, $\bar{\psi}(\mathbf{U}) = \psi\{\bar{s}(\mathbf{U})\}$ and $\widehat{\psi}(\mathbf{U}) = \psi\{\widehat{s}(\mathbf{U})\}$. We assume that $\xi(\cdot)$, the density function of $\bar{\psi}(\mathbf{U})$, is continuously differentiable with bounded derivatives and bounded away from zero on the interval $[\psi(\rho_l), \psi(\rho_r)]$, where $[\rho_l, \rho_r] \subset \Omega_s$ and Ω_s is the support of $\bar{s}(\mathbf{U})$. We also assume that the marker values are bounded, $\boldsymbol{\beta}_k$ belongs to a compact set $\Omega_{\bar{\boldsymbol{\beta}}_k}$. For the bandwidth h, we assume that $h = O(n^{-\nu}), 1/5 < \nu < 1/2$.

A Uniform Consistency of $\widehat{\Delta}(\cdot)$

To derive the asymptotic properties of $\widehat{\Delta}(\cdot)$, we first note that from Tian et al. (2007) and Uno et al. (2007), there exists some deterministic function $\boldsymbol{\Psi}_k$ such that

$$n_k^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}}_k) = n_k^{-\frac{1}{2}} \sum_{i=1}^{n_k} \boldsymbol{\Psi}_k(Y_{ki}, \mathbf{Z}_{ki}) + o_p(1) = O_p(1), \quad \text{for k = 0, 1.} \tag{A·1}$$

Since $\widehat{\Delta}(v) = \widehat{\mu}_1(v) - \widehat{\mu}_0(v) = \mathfrak{g}\{\widehat{a}_1(v)\} - \mathfrak{g}\{\widehat{a}_0(v)\}$, to establish the uniform consistency of $\widehat{\Delta}(v) = \widehat{\mu}_1(v) - \widehat{\mu}_0(v)$, it suffices to show that $\widehat{a}_k(v)$ is uniformly consistent for $a_k(v) = \mathfrak{g}^{-1}\{\bar{\mu}_k(v)\}$,for $k = 0$ and 1. To this end, we aim to show that $\widehat{\mathbf{d}}_k(v) = \{\widehat{d}_{a_k}(v), \widehat{d}_{b_k}(v)\}^\intercal = [\widehat{a}_k(v) - a_k(v), h^{-1}\{\widehat{b}_k(v) - b_k(v)\}]^\intercal \to 0$ in probability uniformly in v, where $b_k(v) = d[\mathfrak{g}^{-1}\{\bar{\mu}_k(v)\}]/dv = \dot{\bar{\mu}}_k(v)/\dot{\mathfrak{g}}\{a_k(v)\}$. At any given v, recall that $\{\widehat{a}_k(v), \widehat{b}_k(v)\}^\intercal$ is the root of the estimating equation (2.3). It follows that $\widehat{\mathbf{d}}_k(v)$ is the solution to the estimating equation

$$\widehat{\mathbb{S}}_k(\mathbf{d}, v) = \begin{bmatrix} \widehat{\mathbb{S}}_{k1}(\mathbf{d}; v) \\ \widehat{\mathbb{S}}_{k2}(\mathbf{d}; v) \end{bmatrix} = n^{-1} \sum_{i=1}^{n} \begin{bmatrix} 1 \\ h^{-1}\widehat{\mathcal{E}}_{kvi} \end{bmatrix} K_h(\widehat{\mathcal{E}}_{kvi}) \left\{ Y_{ki} - \mathcal{G}(\mathbf{d}, v; \widehat{\psi}(\mathbf{U}_{ki}), h) \right\} = 0$$

where $\mathbf{d} = (d_a, d_b)^\intercal$ and $\mathcal{G}(\mathbf{d}, v; y, h) = \mathfrak{g}[a_k(v) + b_k(v)\{y - \psi(v)\} + d_a + d_b h^{-1}\{y - \psi(v)\}]$.

The first step is to show that $\widehat{\mathbb{S}}_k(\mathbf{d}; v)$ is uniformly consistent for

$$\mathbb{S}_k(\mathbf{d}; v) = \begin{bmatrix} \mathbb{S}_{k1}(\mathbf{d}; v) \\ \mathbb{S}_{k2}(\mathbf{d}; v) \end{bmatrix} = \xi(v) \begin{bmatrix} \bar{\mu}_k(v) - \int K(t)\mathfrak{g}\{a_k(v) + d_a + d_b t\}dt \\ -\int tK(t)\mathfrak{g}\{a_k(v) + d_a + d_b t\}dt \end{bmatrix}$$

Since $\sup_{\mathbf{d},v} |\widehat{\mathbb{S}}_{k1}(\mathbf{d}, v) - \mathbb{S}_{k1}(\mathbf{d}, v)| \leq \sup_v |\varepsilon_k^{(1)}(v)| + \sup_{\mathbf{d},v} |\varepsilon_k^{(2)}(\mathbf{d}, v)|$, we first show that

$$\sup_v |\varepsilon_k^{(1)}(v)| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k)\} \quad \text{and} \quad \sup_{\mathbf{d},v} |\varepsilon_k^{(2)}(\mathbf{d}, v)| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k)\}, \qquad \text{(A·2)}$$

$$\text{where} \qquad \varepsilon_k^{(1)}(v) = n_k^{-1} \sum_{i=1}^{n} K_h(\widehat{\mathcal{E}}_{kvi}) Y_{ki} - \xi(v)\bar{\mu}_k(v), \quad \text{and}$$

$$\varepsilon_k^{(2)}(\mathbf{d}, v) = n_k^{-1} \sum_{i=1}^{n} K_h(\widehat{\mathcal{E}}_{kvi}) \mathcal{G}\{\mathbf{d}, v; \widehat{\psi}(\mathbf{U}_{ki}), h)\} - \xi(v) \int K(t)\mathfrak{g}\{a_k(v) + d_a + d_b t\}dt.$$

We only prove the rate of convergence for $\sup_{\mathbf{d},v} |\varepsilon_k^{(2)}(\mathbf{d}, v)|$ since similar arguments could be used to establish the rate of convergence for $\sup_v |\varepsilon_k^{(1)}(v)|$. To this end, we note that from with (A·1) and the same arguments as given in Cai et al. (2008),

$$\left|\widehat{\varepsilon}_k^{(2)}(\mathbf{d}, v)\right| \lesssim n_k^{-\frac{1}{2}} \left(h^{-1} \|\widehat{\mathbb{G}}_k\|_{\mathcal{H}_\epsilon} + \left| \int K_h(y - v) d\widehat{\mathbb{G}}_k \left[\mathcal{G}\{\mathbf{d}, v; \bar{\psi}(\mathbf{U}), h\} I\{\bar{\psi}(\mathbf{U}) \leq y\} \right] \right| \right)$$
$$+ O_p(n_k^{-\frac{1}{2}} + h^2)$$

where $\mathcal{H}_\epsilon = \{I\{g_1(\boldsymbol{\beta}_1'\mathbf{z}) - g_0(\boldsymbol{\beta}_0'\mathbf{z}) \leq y\} - I\{g_1(\bar{\boldsymbol{\beta}}_1'\mathbf{z}) - g_0(\bar{\boldsymbol{\beta}}_0'\mathbf{z}) \leq y\} : \|\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1\| + \|\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}_0\| \leq \epsilon, y\}$ is a class of functions indexed by $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and y. By the maximum inequality of van der Vaart & Wellner (1996) and (A·1), we have $n_k^{-\frac{1}{2}} h^{-1} \|\widehat{\mathbb{G}}_k\|_{\mathcal{H}_\epsilon} \lesssim O_p\{(n_k h)^{-\frac{1}{2}} (n_k h^2)^{-\frac{1}{4}} \log(n_k)\}$. On the other hand, with the standard arguments used in Bickel & Rosenblatt (1973), it can be shown that

$$\left| n_k^{-\frac{1}{2}} \int K_h(y - v) d\widehat{\mathbb{G}}_k \left[\mathcal{G}\{\mathbf{d}, v; \bar{\psi}(\mathbf{U}), h\} I\{\bar{\psi}(\mathbf{U}) \leq y\} \right] \right| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k)\}.$$

Therefore, $\sup_{\mathbf{d},s} |\widehat{\varepsilon}_k^{(2)}(\mathbf{d}, v)| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k)\}$. This implies (A·2) and hence $\sup_{\mathbf{d},v} |\widehat{\mathbb{S}}_{k1}(\mathbf{d}, v) - \mathbb{S}_{k1}(\mathbf{d}, v)| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k)\} = o_p(1)$.

The same arguments as given above can be used to show that $\sup_{\mathbf{d},s} |\widehat{\mathbb{S}}_{k2}(\mathbf{d}, v) - \mathbb{S}_{k2}(\mathbf{d}, v)| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k) + h\} = o_p(1)$. Therefore $\sup_{\mathbf{d},s} |\widehat{\mathbb{S}}_k(\mathbf{d}, v) - \mathbb{S}_k(\mathbf{d}, v)| = o_p(1)$. This uniform convergence, coupled with the fact that $\mathbf{0}$ is the unique solution to the equation $\mathbb{S}_k(\mathbf{d}, v) = 0$ with respect to $\mathbf{d}$ and all the eigenvalues of $\mathbb{A}_k(v) = -\partial \mathbb{S}_k(\mathbf{d}; v)/\partial \mathbf{d}'|_{\mathbf{d}=\mathbf{0}} = \xi(v)\dot{\mathfrak{g}}\{a_k(v)\}\text{diag}\{1, \int v^2 K(v)dv\}$ are uniformly bounded above zero, suggests that $\sup_v |\widehat{\mathbf{d}}_k(v)| = O_p\{(n_k h)^{-\frac{1}{2}} \log(n_k) + h\} = o_p(1)$, which implies the consistency of $\widehat{\mu}_k(v) = \mathfrak{g}\{\widehat{a}_k(v)\}$. Therefore, $\sup_v |\widehat{\Delta}(v) - \bar{\Delta}(v)| \leq \sup_v |\widehat{\mu}_1(v) - \bar{\mu}_1(v)| + \sup_v |\widehat{\mu}_0(v) - \bar{\mu}_0(v)| = o_p(1)$.

B Asymptotic distribution of $\widehat{\mathcal{W}}(v) = (nh)^{\frac{1}{2}}\{\widehat{\Delta}(v) - \bar{\Delta}(v)\}$

It follows from a Taylor series expansion that

$$(n_kh)^{\frac{1}{2}}\widehat{d}_{a_k}(v) = (n_kh)^{\frac{1}{2}}\{\widehat{a}_k(v) - a_k(v)\} = \widehat{\mathbb{B}}_{k1}(v)'(n_kh)^{\frac{1}{2}}\widehat{\mathbb{S}}_k(\mathbf{0}; v) + O_p\left\{(n_kh)^{\frac{1}{2}}(|\widehat{d}_{a_k}(v)|^2 + |\widehat{d}_{b_k}(v)|^2)\right\},$$

where $\widehat{\mathbb{B}}_{k1}(v)$ is the first row of $\widehat{\mathbb{B}}_k(v) = \widehat{\mathbb{A}}_k(v)^{-1}$. Using the similar arguments in the previous section, one can show that $\widehat{\mathbb{B}}_{k1}(v)$ converges to $[\xi(v)^{-1}\dot{\mathfrak{g}}\{a_k(v)\}^{-1}, 0]^{\intercal}$, the first row of $\mathbb{A}_k(v)^{-1}$, uniformly in v. Furthermore, with the convergence rate of $\widehat{\mathbb{S}}_k(\mathbf{d}, v)$, it is not difficult to show that the remainder term is bounded by $O_p\{(n_kh)^{-\frac{1}{2}}\log(n_k)^2 + (n_kh)^{\frac{1}{2}}h^2\}$ uniformly in v. It follows that $(n_kh)^{\frac{1}{2}}\widehat{d}_{a_k}(v) = \frac{(n_kh)^{\frac{1}{2}}\widehat{\mathbb{S}}_{k1}(\mathbf{0};v)}{\xi(v)\dot{\mathfrak{g}}\{a_k(v)\}} + O_p\{(n_kh)^{-\frac{1}{2}}\log(n_k)^2 + (n_kh)^{\frac{1}{2}}h^2\}$. Coupled with the convergence rate of $\widehat{\mathbb{S}}_{k1}(\mathbf{0}; v)$, we have $\sup_v |\widehat{d}_{a_k}(v)| = O_p\{(n_kh)^{-\frac{1}{2}}\log(n_k)\}$. Thus,

$$(n_kh)^{\frac{1}{2}}\widehat{d}_{a_k}(v) = [\xi(v)\dot{\mathfrak{g}}\{a_k(v)\}]^{-1}(n_kh)^{\frac{1}{2}}\widehat{\mathbb{P}}_k\left[K_h(\widehat{\mathcal{E}}_{kv})\{Y_k - \eta_k(v, \widehat{\psi}(\mathbf{U}_k))\}\right] + o_p(1).$$

where $\eta_k(v, y) = \mathfrak{g}[a_k(v) + b_k(v)\{y - \psi(v)\}]$. We next show that

$$(n_kh)^{\frac{1}{2}}\widehat{d}_{a_k}(v) = (n_kh)^{\frac{1}{2}}\widetilde{d}_{a_k}(v) + o_p(1), \tag{B·1}$$

where $(n_kh)^{\frac{1}{2}}\widetilde{d}_{a_k}(v) = [\xi(v)\dot{\mathfrak{g}}\{a_k(v)\}]^{-1}(n_kh)^{\frac{1}{2}}\widehat{\mathbb{P}}_k\left(K_h(\mathcal{E}_{kv})\left[Y_k - \eta_k\{v, \bar{\psi}(\mathbf{U}_k)\}\right]\right)$ and $\mathcal{E}_{kv} = \bar{\psi}(\mathbf{U}_k) - \psi(v)$. Noticing the fact that $\xi(v)\dot{\mathfrak{g}}\{a_k(v)\}$ is bounded away from zero uniformly in v, we have

$$(n_kh)^{\frac{1}{2}}\left|\widehat{d}_{a_k}(v) - \widetilde{d}_{a_k}(v)\right| \lesssim h^{-\frac{1}{2}}\|\widehat{\mathbb{G}}_k\|_{\mathcal{F}_\epsilon} + h^{-\frac{1}{2}}\|\widehat{\mathbb{G}}_k\|_{\mathcal{H}_\epsilon} + O_p\{(n_kh)^{\frac{1}{2}}|\widehat{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1| + (n_kh)^{\frac{1}{2}}|\widehat{\boldsymbol{\beta}}_0 - \bar{\boldsymbol{\beta}}_0| + h^2\}$$

where $\mathcal{F}_\epsilon = \left\{yI\{g_1(\boldsymbol{\beta}_1^{\intercal}\mathbf{z}) - g_0(\boldsymbol{\beta}_0^{\intercal}\mathbf{z}) \le c\} - yI\{g_1(\bar{\boldsymbol{\beta}}_1^{\intercal}\mathbf{z}) - g_0(\bar{\boldsymbol{\beta}}_0^{\intercal}\mathbf{z}) \le c\} : \|\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1\| + \|\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}_0| \le \epsilon, c\right\}$ is the class of functions indexed by $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and c. By the maximum inequality and (A·1) we have $h^{-\frac{1}{2}}\|\widehat{\mathbb{G}}_k\|_{\mathcal{F}_\epsilon} = O_p\{h^{-\frac{1}{2}}n^{-\frac{1}{4}}\log(n_k)\}$. This along with the convergence rate for $h^{-\frac{1}{2}}\|\widehat{\mathbb{G}}_k\|_{\mathcal{H}_\epsilon}$ implies (B·1). Furthermore, by a delta method and the standard arguments for local linear regression fitting, we have

$$\widehat{\mathcal{W}}(v) \simeq (nh)^{\frac{1}{2}}\{\xi(v)\}^{-1}\left(\widehat{\mathbb{P}}_1\left[K_h(\mathcal{E}_{1v})\left\{Y_1 - \bar{\mu}_1(v)\right\}\right] - \widehat{\mathbb{P}}_0\left[K_h(\mathcal{E}_{0v})\left\{Y_0 - \bar{\mu}_0(v)\right\}\right]\right) \tag{B·2}$$

which converges to a normal with mean 0 and variance $\sigma_1(v)^2 + \sigma_0(v)^2$, where

$$\sigma_k^2(v) = m_2\{p_k\xi(v)^2\}^{-1}\mathrm{var}\{Y_k|\bar{s}(\mathbf{U}_k) = v\} \quad \text{and} \quad m_2 = \int K(v)^2dv.$$

To justify the resampling method, we first note that since $|\widehat{\boldsymbol{\beta}}_1^* - \widehat{\boldsymbol{\beta}}_1| + |\widehat{\boldsymbol{\beta}}_0^* - \widehat{\boldsymbol{\beta}}_0| = O_p(n^{-\frac{1}{2}})$, one may use the same argument as given above to show that

$$\widehat{\mathcal{W}}^*(v) = (nh)^{\frac{1}{2}} \left[\frac{\sum_{i=1}^{n_1} K_h(\widehat{\mathcal{E}}_{1vi})\{Y_{1i} - \widehat{\mu}_1(v)\}\mathcal{Z}_{1i}}{\sum_{i=1}^{n_1} K_h(\widehat{\mathcal{E}}_{1vi})} - \frac{\sum_{i=1}^{n_0} K_h(\widehat{\mathcal{E}}_{0vj})\{Y_{0j} - \widehat{\mu}_0(v)\}\mathcal{Z}_{0j}}{\sum_{j=1}^{n_0} K_h(\widehat{\mathcal{E}}_{0vj})} \right] + o_p(1).$$

Furthermore, conditional on the observed data, $(nh)^{\frac{1}{2}}\widehat{\mathbb{P}}_k[K_h(\widehat{\mathcal{E}}_{kv})\{Y_k - \widehat{\mu}_k(v)\}\mathcal{Z}]/\widehat{\mathbb{P}}_k\{K_h(\widehat{\mathcal{E}}_{kv})\}$ is asymptotical normally distributed with mean 0 and variance

$$\widehat{\sigma}_k^2(v) = h\widehat{\mathbb{P}}_k[K_h(\widehat{\mathcal{E}}_{kv})^2\{Y_k - \widehat{\mu}_k(v)\}^2]/[p_k\widehat{\mathbb{P}}_k\{K_h(\widehat{\mathcal{E}}_{kv})\}]^2.$$

It follows from the arguments given in Appendix A to show that $\widehat{\sigma}_k^2(v)$ converges to $\sigma_k^2(v)$, as $n \to \infty$.

C Justification for the Validity of the Confidence Band for $\bar{\Delta}(v)$

We first justify that after proper standardization, the supermum type statistics $\widehat{\mathcal{M}}$ converges weakly. It follows from (B·2) and the consistency of $\widehat{\sigma}(v)$ for $\sigma(v)$ that

$$\widehat{\mathcal{M}} = \sup_v \left| (nh)^{\frac{1}{2}} \frac{\widehat{\mathbb{P}}_1\left[K_h(\mathcal{E}_{1v})\left\{Y_1 - \bar{\mu}_1(v)\right\}\right] - \widehat{\mathbb{P}}_0\left[K_h(\mathcal{E}_{0v})\left\{Y_0 - \bar{\mu}_0(v)\right\}\right]}{\xi(v)\sigma(v)} \right| + o_p(1).$$

This, together with the continuity of $\bar{\mu}_k(\cdot)$ and $\xi(v)\sigma(v)$, implies that

$$\widehat{\mathcal{M}} = \sup_v \left| n^{-\frac{1}{2}} h^{\frac{1}{2}} \sum_{j=1}^{n} K_h(\mathcal{E}_{vj})\mathcal{V}_j \right| + o_p(1).$$

where we rewrite the data as $\{(Y_j, \mathbf{U}_j, G_j), j = 1, ..., n\}$ with G_j being the treatment group indicator for the jth subject, $\mathcal{E}_{vj} = \bar{\psi}(\mathbf{U}_j) - \psi(v)$, and

$$\mathcal{V}_j = (-1)^{G_j+1} \frac{Y_j - G_j\bar{\mu}_1\{\bar{s}(\mathbf{U}_j)\} - (1-G_j)\bar{\mu}_0\{\bar{s}(\mathbf{U}_j)\}}{\xi\{\bar{s}(\mathbf{U}_j)\}\sigma\{\bar{s}(\mathbf{U}_j)\}}$$

Using similar argument in Bickel & Rosenblatt (1973), we have $\mathrm{pr}\{a_n(\widehat{\mathcal{M}} - d_n) < x\} \to e^{-2e^{-x}}$, where

$$a_n = (2\log[\{\psi(\rho_r) - \psi(\rho_l)\}/h])^{\frac{1}{2}} \quad \text{and} \quad d_n = a_n + a_n^{-1}\log\left\{\int \dot{K}(t)^2 dt/(4m_2\pi)\right\}.$$

Now, to justify the resampling procedure for constructing the confidence band, we note that

$$\widehat{\mathcal{M}}^* = \sup_v \left| n^{-\frac{1}{2}} h^{\frac{1}{2}} \sum_{j=1}^{n} K_h(\widehat{\mathcal{E}}_{vj})\widehat{V}_j Z_j + (nh)^{\frac{1}{2}} \left\{\widehat{\Delta}(v; \widehat{\boldsymbol{\beta}}_1^*, \widehat{\boldsymbol{\beta}}_0^*) - \widehat{\Delta}(v)\right\} \right|,$$

where $\widehat{V}_j$ is obtained by replacing all the theoretical quantities in $\mathcal{V}_j$ by their empirical counterparts. Again, since $|\widehat{\boldsymbol{\beta}}_1^* - \widehat{\boldsymbol{\beta}}_1| + |\widehat{\boldsymbol{\beta}}_0^* - \widehat{\boldsymbol{\beta}}_0| = O_p(n^{-\frac{1}{2}})$, from Appendix B, we have

$$\widehat{\mathcal{M}}^* = \sup_v \left| n^{-\frac{1}{2}} h^{\frac{1}{2}} \sum_{j=1}^{n} K_h(\widehat{\mathcal{E}}_{vj}) \widehat{V}_j Z_j \right| + o_p(1).$$

It follows from the same argument as given in Tian et al. (2005) that,

$$\sup_x \left| \text{pr} \left\{ a_n(\widehat{\mathcal{M}}^* - d_n) < x \mid (Y_i, \mathbf{U}_i, G_i), i = 1, \cdots, n \right\} - e^{-2e^{-x}} \right| \to 0$$

in probability as $n \to \infty$. Therefore, the conditional distribution of $a_n(\widehat{\mathcal{M}}^* - d_n)$ can be used to approximate the distribution of $a_n(\widehat{\mathcal{M}} - d_n)$ for large n.

References

Bickel, P. J. & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates (Corr: V3 p1370). *The Annals of Statistics* **1**, 1071–1095.

Bonetti, M. & Gelber, R. D. (2000). A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. *Statistics in Medicine* **19**, 2595–609.

Bonetti, M. & Gelber, R. D. (2005). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5**, 465–81.

Bowman, A., Hall, P. & Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika* **85**, 799–808.

Byar, D. P. (1985). Assessing apparent treatment: Covariate interactions in randomized clinical trials. *Statistics in Medicine* **4**, 255–263.

Cai, T., Tian, L., Lloyd-Jones, D. & Wei, L. J. (2008). Evaluating subject-level incremental values of new markers for risk classification rule. *Harvard Biostatistics Working Paper Series* .

Fan, J. & Gijbels, I. (1996). Local polynomial modelling and its applications, Vol. 66 of Monographs on Statistics and Applied Probability. *London: Chapman Hall,* .

Hammer, S., Squires, K., Hughes, M., Grimes, J., Demeter, L., Currier, J., Eron, J., Feinberg, J. Balfour, H., Deyton, L., Chodakewitz, J., Fischl, M., Phair, J., Spreen, W., Pedneault, L., Nguyen, B., Cook, J. & ACTG 320 Study Team (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *N. Engl. J. Med.* **337**, 725–33.

IBCSG (2002). (The International Breast Cancer Study Group) endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node negative breast cancer: A randomized trial. *J Natl Cancer Inst* **94**, 1054–65.

Park, B., Kim, W., Ruppert, D., Jones, M., Signorini, D. & Kohn, R. (1997). Simple transformation techniques for improved non-parametric regression. *Scandinavian journal of statistics* **24**, 145–163.

Park, Y. (2004). *Semiparametric Statistical Inference in Survival Analysis*. PhD thesis, Harvard University.

Pfeffer, M. & Jarcho, J. (2006). The Charisma of Subgroups and the Subgroups of CHARISMA. *New England Journal of Medicine* **354**, 1744.

Rothwell, P. (2005). External validity of randomised controlled trials:To whom do the results of this trial apply?. *The Lancet* **365**, 82–93.

Sabine, C. (2005). AIDS events among individuals initiating HAART: do some patients experience a greater benefit from HAART than others?. *AIDS* **19**, 1995.

Song, X. & Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60**, 874–83.

Tian, L., Cai, T., Goetghebeur, E. & Wei, L. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**, 297.

Tian, L., Zucker, D. & Wei, L. (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association* **100**, 172–183.

Tibshirani, R. & Hastie, T. (1984). *Local likelihood estimation.* SLAC-275, Stanford Linear Accelerator Center, CA (USA).

Uno, H., Cai, T., Tian, L. & Wei, L. (2007). Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models. *Journal of the American Statistical Association* **102**, 527.

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

Wand, M., Marron, J. & Ruppert, D. (1991). Transformation in density estimation (with comments). *Journal of the American Statistical Association* **86**, 343–361.

Wang, R., Lagakos, S., Ware, J., Hunter, D. & Drazen, J. (2007). Statistics in Medicine–Reporting of Subgroup Analyses in Clinical Trials. *New England Journal of Medicine* **357**, 2189.

Table 1: Estimated regression coefficients and their standard error estimates for the two treatment groups with the data from ACTG 320.

(a) Continuous outcome: change in CD4 from baseline to week 24

		$\log_{10}$ RNA	CD4	log(CD4)	Age
Two Drug	Estimate	3.47	0.04	-0.07	0.44
	Std. Error	3.40	0.07	3.91	0.26
Three Drug	Estimate	28.65	-0.24	24.14	-0.22
	Std. Error	6.65	0.14	8.05	0.49

(b) Binary outcome: RNA at week 24 $\leq$ 500 copies/ml

		$\log_{10}$ RNA	log(CD4)	Age
Two Drug	Estimate	-1.14	-0.31	0.00
	Std. Error	0.22	0.16	0.02
Three Drug	Estimate	-0.60	0.32	0.06
	Std. Error	0.18	0.10	0.01

Figure 1: (a): The estimated density function of the parametric score with respect to Week 24 CD4 changes; (b): Estimated group averages of Week 24 CD4 changes over the score for two- and three-drug combination groups; (c): Estimated treatment differences (thick curve), three drug combo minus two drug combo, with respect to Week 24 CD4 changes over the score, and the corresponding 95% pointwise (dashed curve) and simultaneous (shaded region) confidence intervals.

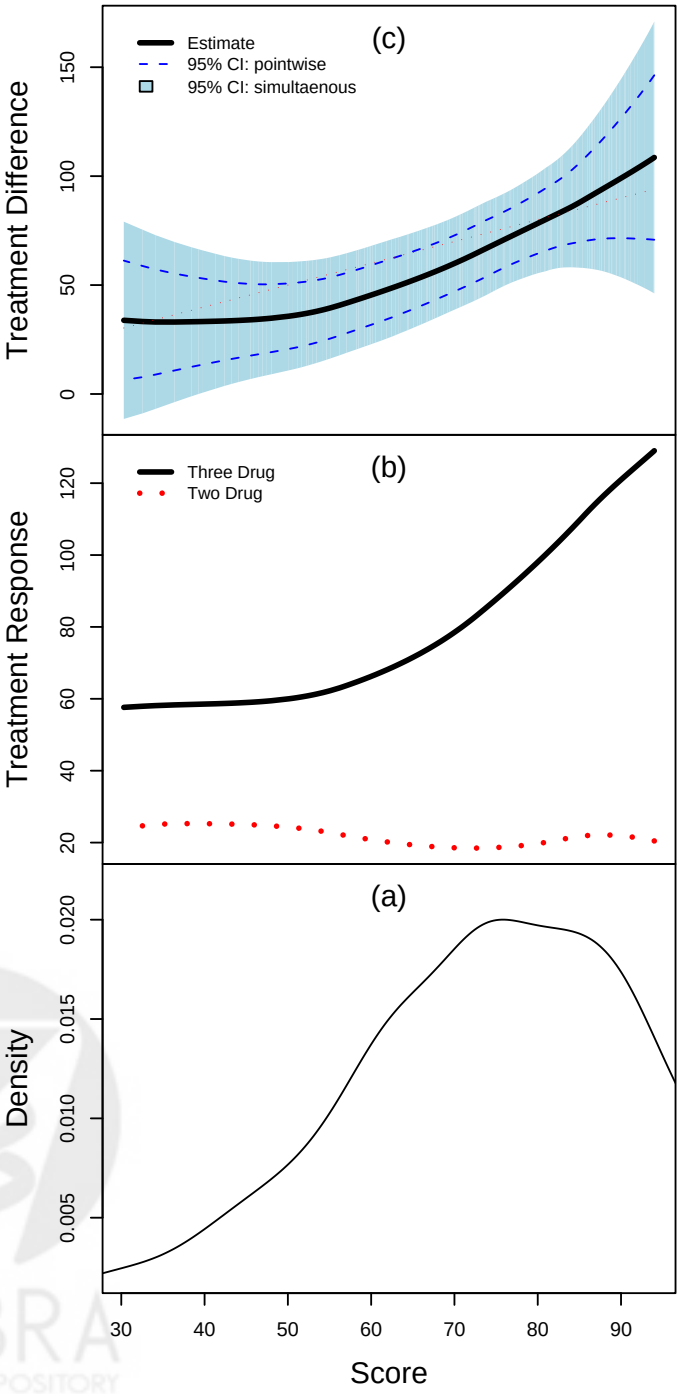

Figure 2: (a): The estimated density function of the parametric score with respect to Week 24 HIV-RNA; (b): Estimated averages of Week 24 RNA over the score for two- and three-drug combination groups; (c): Estimated treatment differences (thick curve), three drug combo minus two drug combo, with respect to Week 24 RNA over the score, and the corresponding 95% pointwise (dashed curve) and simultaneous (shaded region) confidence intervals.