Harvard University
Harvard University Biostatistics Working Paper Series

Year 2011 *Paper* 134

Effectively Selecting a Target Population for a Future Comparative Study

Lihui Zhao* Lu Tian† Tianxi Cai‡
Brian Claggett** L. J. Wei††

*Northwestern University, lihui.zhao@northwestern.edu
†Stanford University School of Medicine, lutian@stanford.edu
‡Harvard University, tcai@hsph.harvard.edu
**Harvard University, bclagget@hsph.harvard.edu
††Harvard University, wei@hsph.harvard.edu

Effectively Selecting a Target Population for a Future Comparative Study

Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, and L. J. Wei

Abstract

When comparing a new treatment with a control in a randomized clinical study, the treatment effect is generally assessed by evaluating a summary measure over a specific study population. The success of the trial heavily depends on the choice of such a population. In this paper, we show a systematic, effective way to identify a promising population, for which the new treatment is expected to have a desired benefit, using the data from a current study involving similar comparator treatments. Specifically, with the existing data we first create a parametric scoring system using multiple covariates to estimate subject-specific treatment differences. Using this system, we specify a desired level of treatment difference and create a subgroup of patients, defined as those whose estimated scores exceed this threshold. An empirically calibrated group-specific treatment difference curve across a range of threshold values is constructed. The population of patients with any desired level of treatment benefit can then be identified accordingly. To avoid any "self-serving" bias, we utilize a cross-training-evaluation method for implementing the above two-step procedure. Lastly, we show how to select the best scoring system among all competing models. The proposals are illustrated with the data from two clinical trials in treating AIDS and cardiovascular diseases. Note that if we are not interested in designing a new study for comparing similar treatments, the new procedure can also be quite useful for the management of future patients who would receive nontrivial benefits to compensate for the risk or cost of the new treatment.

EFFECTIVELY SELECTING A TARGET POPULATION FOR A FUTURE COMPARATIVE STUDY

Lihui Zhao[1], Lu Tian[2], Tianxi Cai[3], Brian Claggett[3], and L. J. Wei[3]*

ABSTRACT

When comparing a new treatment with a control in a randomized clinical study, the treatment effect is generally assessed by evaluating a summary measure over a specific study population. The success of the trial heavily depends on the choice of such a population. In this paper, we show a systematic, effective way to identify a promising population, for which the new treatment is expected to have a desired benefit, using the data from a current study involving similar comparator treatments. Specifically, with the existing data we first create a parametric scoring system using multiple covariates to estimate subject-specific treatment differences. Using this system, we specify a desired level of treatment difference and create a subgroup of patients, defined as those whose estimated scores exceed this threshold. An empirically calibrated group-specific treatment difference curve across a range of threshold values is constructed. The population of patients with any desired level of treatment benefit can then be identified accordingly. To avoid any "self-serving" bias, we utilize a cross-training-evaluation method for implementing the above two-step procedure. Lastly, we show how to select the best scoring system among all competing models. The proposals are illustrated with the data from two clinical trials in treating AIDS and cardiovascular diseases. Note that if we are not interested in designing a new study for comparing similar treatments, the new procedure can also be quite useful for the management of future patients who would receive nontrivial benefits to compensate for the risk or cost of the new treatment.

Keywords: Cross-training-evaluation; Lasso procedure; Personalized medicine; Prediction; Ridge regression; Stratified medicine; Subgroup analysis; Variable selection.

[1]Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA
[2]Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA
[3]Department of Biostatistics, Harvard University, Boston, MA 02115, USA

1. INTRODUCTION

In comparing a new treatment with a control via a randomized clinical trial, the assessment of the treatment efficacy is usually based on an overall summary measure over a specific study population. To increase the chance of success of the study, it is important to choose an appropriate study population for which the new treatment is expected to have non-trivial overall benefits that compensate for its risks and/or costs. In this paper we are interested in developing strategies which identify such a patient population with the data from a current study for comparing similar treatments. Even when we are not interested in designing another new study for comparing similar treatments, the new proposal provides a systematic, efficient procedure for management of future patients with the new treatment.

As an example, one of the very first trials to evaluate the added value of a potent protease inhibitor, indinavir, for HIV patients, was conducted by the AIDS Clinical Trials Group (ACTG). This randomized, double-blind study, ACTG 320 (Hammer et al., 1997), compared a three-drug combination, indinavir, zidovudine and lamivudine, with the standard two-drug combination, zidovudine and lamivudine. There were 1156 patients enrolled in the study. One of the endpoints was the CD4 count, measured 24 weeks after randomization. The overall estimated mean difference between the new treatment and control over the entire study population was 81 cells/mm^3. Although the overall efficacy from the three-drug combination group is highly statistically significant, it is not necessarily true that the new therapy works for all future patients. Moreover, there are nontrival toxicities and serious concerns about the development of protease inhibitor resistance mutations. Now, suppose that having an expected treatment benefit representing a week 24 CD4 count increase of 100 cells/mm^3 relative to the control would be sufficient to compensate for the costs and risks of using the new therapy. The question, then, is how to identify such a subpopulation efficiently via the "baseline" covariates.

Various novel quantitative methods have been proposed to deal with the problem of heterogenous treatment effects. For cases with a single covariate, Song and Pepe (2004), assuming a monotone relationship between the covariate and the treatment difference, proposed a procedure to obtain an optimal division of the population for determining which future patients should receive the treatment or control. Song and Zhou (2011) generalized this method for censored event time data. Janes et al. (2011) gave a practical guidance on using the marker-by-treatment predictiveness curves for treatment selection. Moreover, Bonetti and Gelber (2000, 2005) stratified patients utilizing a moving average procedure to obtain subject-specific nonparametric estimates for the treatment difference. For cases with multiple covariates, Cai et al. (2010) proposed a systematic two stage method for personalized treatment selection using a parametric scoring system for estimating the subject-specific treatment difference, followed by a nonparametric smoothing technique at the second stage. However, it is not clear how to utilize their procedure to efficiently identify a group of future patients who would have a desired overall treatment benefit. Moreover, there are no procedures available in the literature for comparing different scoring systems for treatment differences with multiple covariates.

Note that if the scoring system is built using data from the control group only, one may not be able to effectively identify a target population which has a desirable overall treatment benefit. For example, high-risk patients do not necessarily experience the greatest benefit from a new treatment. Thus for the present problem, even when considering only a single covariate, a first step is to create a scoring system for estimating the *treatment difference*; one can then use this system to identify such a target population effectively. However, unlike the prediction problem in the one sample case, none of the existing procedures in the literature which use scores for estimating treatment differences can be utilized to directly evaluate the performance of competing scoring systems. This difficulty arises from the fact that each study subject was assigned to receive either the new treatment or control, but not both. Therefore, it is not clear how to compare, at the patient level, the observed treatment difference to its predicted counterpart.

For the case of a single treatment group, Moskowitz and Pepe (2004) generalize the idea of positive predictive values (PPV) and negative predictive values to accommodate a single continuous covariate and a binary outcome, and propose a graphical method to summarize predictive accuracy. In this paper, we generalize the notion of PPV to handle the present problem of treatment selections with multiple baseline covariates. Specifically, we first generate various parametric scoring systems for estimating the subject-specific treatment differences using baseline markers, and select the "best" one among all the candidate models. Various criteria are utilized for model selection based on, for example, the concordance between the observed and expected treatment differences. We then show how to define a target patient population, which can be used for identifying future patients who would benefit from the new treatment for the purpose of designing inclusion/exclusion criteria for enrollment in future clinical trials. Our procedure does not require the usage of nonparametric smoothing techniques, which can be quite unstable when the sample size is not large. We illustrate our methods using the data from the above HIV study and also the censored survival time data from a large cardiovascular trial to compare the efficacy of Angiotensin-converting-enzyme inhibitors (ACEi) to a conventional therapy for patients with stable coronary heart disease and preserved left ventricular function (Braunwald et al., 2004).

2. SELECTING THE TARGET SUBPOPULATION VIA A SCORING SYSTEM

Suppose that each subject in a comparative study was randomly assigned to one of the two groups, denoted by $G = 0$ (control) or 1 (treatment). Let $\pi_k = \mathrm{pr}(G = k)$ for $k = 0, 1$. Let Z be the patient's p-dimensional vector of baseline covariates, and $Y_{(k)}$ be the response variable or a function thereof, if the subject had been assigned to Group k, $k = 0, 1$. For each subject, only $Y = GY_{(1)} + (1 - G)Y_{(0)}$ can potentially be observed. Assume that a larger Y indicates a better clinical outcome. For ease of presentation, we first consider the non-censored case that for each subject, we can observe the triplet (Y, G, Z) completely.

Now, let $\mu_k(Z) = E(Y_{(k)}|Z)$ be the expected response for patients in Group k, conditional on Z. Furthermore, let the treatment difference $D(Z) = \mu_1(Z) - \mu_0(Z)$. The data,

$\{(Y_i, G_i, Z_i); i = 1, \ldots, n\}$, consist of n independent copies of (Y, G, Z). Suppose that $\hat{D}(Z)$ is an estimator for $D(Z)$. Let Z^0 be the covariate vector for a future patient randomly drawn from the same population of the current study. Also let $Y^0_{(k)}$ be the potential response of this patient if assigned to Group k, $k = 0, 1$. Consider the subgroup of subjects such that $\hat{D}(Z^0) \geq c$, where c is some fixed constant. That is, this subgroup of subjects has an estimated treatment difference no less than c. Let $AD(c)$ be the average treatment difference for this subgroup of subjects:

$$E\left(\left(Y^0_{(1)} - Y^0_{(0)}\right) | \hat{D}(Z^0) \geq c\right), \tag{2.1}$$

where the expectation is with respect to $Y^0_{(k)}$ and Z^0, and also $\{(Y_i, G_i, Z_i); i = 1, \ldots, n\}$. The $AD(c)$ can be estimated by

$$\hat{AD}(c) = \frac{\sum_i^n Y_i I\{\hat{D}(Z_i) \geq c, G_i = 1\}}{\sum_i^n I\{\hat{D}(Z_i) \geq c, G_i = 1\}} - \frac{\sum_i^n Y_i I\{\hat{D}(Z_i) \geq c, G_i = 0\}}{\sum_i^n I\{\hat{D}(Z_i) \geq c, G_i = 0\}}, \tag{2.2}$$

where $I(\cdot)$ is the indicator function. Note that $\hat{AD}(c)$ may not be stable when c is in the upper tail of the distribution of $\hat{D}(Z^0)$.

As a function of c, $\hat{AD}(c)$ can be quite useful for identifying patients who can expect specific levels of benefit from the new treatment relative to the control. As an example, consider the ACTG 320 study discussed in the Introduction. For simplicity, let Y be the CD4 count at week 24 and Z be a vector consisting of two baseline covariates, log(CD4) and $\log_{10}$(RNA). These two measures have been shown to be highly predictive of various clinical outcomes relevant to HIV disease. One may obtain $\hat{D}(Z)$ by the difference of two estimates $\hat{\mu}_0(Z)$ and $\hat{\mu}_1(Z)$ based on two separate additive linear regression models, as given in Table 1. The resulting score for estimating the treatment difference is given by

$$\hat{D}(Z) = -120.61 + 12.57 \log(\text{CD4}) + 29.13 \log_{10}(\text{RNA}).$$

Note that a patient with high baseline CD4 and RNA values is expected to benefit more from the new treatment. Figure 1 provides the estimated $\hat{AD}(c)$ in (2.2) over a range of values c. As discussed in the Introduction, the new treatment, a three-drug combination, has an impressive overall efficacy benefit with regards to week 24 CD4 count. At the time, there were concerns about the cost of the new therapy, as well as the potential for toxicity and/or development of drug resistance. Suppose that, in order to compensate for such non-trivial risks, one would like to treat future patients whose anticipated benefit from the new treatment, relative to the two-drug combination, is "clinically" significant. For example, a meaningful benefit may be defined as an overall CD4 count difference, between the two treatments, of 100 cells/mm^3 at week 24. From Figure 1, $\hat{AD}(77) = 100$, thus this subset of patients would be composed of patients with Z^0 such that $\hat{D}(Z^0) \geq 77$.

Now, let us consider the case that the response variable may not be observed completely. For instance, let T be an event time and $Y = I(T \geq t_0)$, where t_0 is a specific time point of interest. Often T may be censored by a censoring variable C, which is assumed to be independent of T and Z given G. For each subject, the observed quantity is (X, Δ, G, Z), where $X = \min(T, C)$ and $\Delta = I(T \leq C)$. The data, $\{(X_i, \Delta_i, G_i, Z_i); i = 1, \ldots, n\}$, consist of n independent copies of (X, Δ, G, Z). For this case, the $AD(c)$ can be estimated by the difference in Kaplan-Meier survival probabilities, i.e.,

$$\hat{AD}(c) = \prod_{t=0}^{t_0} \left\{ 1 - \frac{\sum_{i=1}^n dN_{i,c}^{(1)}(t)}{\sum_{i=1}^n Y_{i,c}^{(1)}(t)} \right\} - \prod_{t=0}^{t_0} \left\{ 1 - \frac{\sum_{i=1}^n dN_{i,c}^{(0)}(t)}{\sum_{i=1}^n Y_{i,c}^{(0)}(t)} \right\}, \tag{2.3}$$

where $N_{i,c}^{(k)}(t) = I(X_i \leq t, \hat{D}(Z_i) \geq c, G_i = k)\Delta_i$, and $Y_{i,c}^{(k)}(t) = I(X_i \geq t, \hat{D}(Z_i) \geq c, G_i = k)$, $k = 0, 1$; $i = 1, \ldots, n$. Note that $\prod$ here denotes a product integral operator.

If one is interested in a global treatment contrast measure rather than t_0-year survival rates, the standard hazard ratio estimate may be utilized for building a scoring system. When the proportional hazards assumption is violated, however, it is not clear which parameter this model-based estimate would converge to (Prentice and Kalbfleisch, 1981; Lin and Wei, 1989, Xu and O'Quigley, 2000). The overall mean survival time is generally difficult to estimate well due to censoring. On the other hand, one may consider the restricted mean survival time up to a specific time point (Irwin, 1949; Andersen et al., 2004), say, τ_0, as an overall measure for quantifying survivorship. Note that this mean value is simply the area under the corresponding Kaplan-Meier curve truncated at time τ_0. To this end, for the present problem, we let $Y = \min(T, \tau_0)$. It is straightforward to show that the corresponding $AD(c)$ can be estimated by

$$\hat{AD}(c) = \int_0^{\tau_0} \left[\prod_{s=0}^{t} \left\{ 1 - \frac{\sum_{i=1}^n dN_{i,c}^{(1)}(s)}{\sum_{i=1}^n Y_{i,c}^{(1)}(s)} \right\} \right] dt - \int_0^{\tau_0} \left[\prod_{s=0}^{t} \left\{ 1 - \frac{\sum_{i=1}^n dN_{i,c}^{(0)}(s)}{\sum_{i=1}^n Y_{i,c}^{(0)}(s)} \right\} \right] dt, \quad (2.4)$$

using the fact that $E\{\min(T, \tau_0) | \hat{D}(Z^0) \geq c\} = \int_0^{\tau_0} \text{pr}(T > t | \hat{D}(Z^0) \geq c) dt$.

Now, suppose that the covariate vector $Z^0 \in \Omega$ is bounded. In addition, we assume that $\hat{D}(Z^0)$ converges in probability to a finite constant $\bar{D}(Z^0)$ uniformly in $Z^0 \in \Omega$, as $n \to \infty$. Note that $\bar{D}(Z^0)$ could be different from $D(Z^0)$ when the working model is misspecified. Let $\bar{AD}(c) = E\left((Y_{(1)}^0 - Y_{(0)}^0) | \bar{D}(Z^0) \geq c \right)$. In Appendix A, we show that

$$\sup_{c \in (-\infty, c_0)} |AD(c) - \bar{AD}(c)| = o(1), \quad (2.5)$$

for any c_0 such that $\text{pr}\left(\bar{D}(Z^0) \geq c_0 \right) > 0$. Furthermore, for $\hat{AD}(c)$ defined in (2.2) to (2.4),

$$\sup_{c \in (-\infty, c_0)} |\hat{AD}(c) - \bar{AD}(c)| = o_p(1), \quad (2.6)$$

i.e., $\hat{AD}(c)$ is uniformly consistent for $\bar{AD}(c)$, for $c \in (-\infty, c_0)$.

Given a particular scoring system, a plot like Figure 1 is useful for identifying the target patient population who would benefit from the new treatment at various levels of interest. However, it is likely there are other scoring systems using baseline variables which could be better than the present one. In the next section, we discuss how to compare different scoring systems.

3. COMPARING SCORING SYSTEMS

For a reasonably good scoring system, one expects that the curve $\hat{AD}(c)$ is increasing over c, as in Figure 1. In general, different scoring systems $\hat{D}(\cdot)$ will group patients differently. In order to compare two systems, say $\hat{D}_1(\cdot)$ and $\hat{D}_2(\cdot)$, we need to modify the scale of the x-axis for the plot in Figure 1. Specifically, we convert the conditional event $\hat{D}(Z^0) \geq c$ in (2.1) to $H(\hat{D}(Z^0)) \geq q$, where H is the empirical cumulative distribution function of $\hat{D}(Z^0)$. The resulting estimate corresponding to (2.2) is denoted by $\tilde{AD}(q)$. Note that $\tilde{AD}(q) = \hat{AD}(H^{-1}(q))$. Given $0 \leq q \leq 1$, $\tilde{AD}(q)$ is simply an estimated average treatment difference for subjects with scores exceeding the qth quantile. For example, with this new scale for the x-axis, the curve in Figure 1 becomes the solid curve $\tilde{AD}_1(q)$ in Figure 2. The subgroup of patients with an average CD4 count treatment difference of 100, as described in Section 2, represents the patients with scores in the top 52%. Now, since RNA is relatively expensive to measure in resource-limited regions, one question is whether we can use the baseline log(CD4) only to construct a scoring system $\hat{D}_2(\cdot)$ (see Table 2). The resulting score is $\hat{D}_2(Z) = 40.57 + 8.27\log(\text{CD4})$. Note that this new score indicates that a patient with a large baseline CD4 value tends to benefit more from the new treatment. The corresponding $\tilde{AD}_2(q)$ is given in Figure 2 (dashed curve). This new curve is not an increasing function. Moreover, this curve is uniformly lower than $\tilde{AD}_1(q)$, indicating that the addition of baseline RNA allows for substantial improvement in selecting the subgroup of patients with a desirable level of overall treatment benefit. In general, the higher the curve $\tilde{AD}(\cdot)$ is, the better the scoring system is. It is interesting to note that if we were able to use the score $\hat{D}(Z) = D(Z)$, the true treatment difference, the resulting curve $\tilde{AD}(\cdot)$ would be uniformly the best among all working models for treatment differences based on Z (see Appendix B for details). However, when the dimension of Z is greater than one, it is difficult, if not impossible, to estimate $D(Z)$ well nonparametrically.

It is likely that the treatment difference curve $\tilde{AD}(\cdot)$ resulting from one model may not dominate that from another model over the entire interval of interest. If we are interested in identifying a subpopulation with a specific treatment difference, one may choose a model which gives us the largest subset of patients satisfying this criteria among all candidate models. If there is no specific level of treatment difference that is of particular clinical interest, one may use a summary of the curve to select the "best" model. For example, a possible metric is the area under the curve (AUC) of $\tilde{AD}(\cdot)$. Let $\bar{H}(\cdot)$ denote the cumulative distribution function of $\bar{D}(Z^0)$. In Appendix C, we show that the AUC is a consistent estimator for

$$E\left(D(Z^0)\log\left\{\left[1-\bar{H}(\bar{D}(Z^0))\right]^{-1}\right\}\right), \tag{3.1}$$

which is the expected value of the product of the true subject-specific treatment difference $D(Z^0)$, given the individual patient's covariate vector Z^0, and a strictly increasing transformation of the rank of the patient's limiting score $\bar{D}(Z^0)$. The quantity (3.1) is a measure of the concordance between the true treatment difference and its empirical estimate. Therefore, a higher AUC indicates a better fit of the working model. Furthermore, the area between the curves (ABC) of $\tilde{AD}(\cdot)$ and the horizontal line $y = \tilde{AD}(0)$, estimates the corresponding covariance of two random quantities in (3.1). Note that this covariance is $\rho\sigma_0$, where ρ is the correlation of the two terms in (3.1) and σ_0 is an unknown constant which does not depend on any specific scoring system. It follows that to compare two scoring systems, one may use the ratio of two ABCs to examine the relative improvement from one model to the other.

Since the tail part of the curve $\tilde{AD}(\cdot)$ may not be stable, one may use a partial AUC (by integrating the curve up to a specific constant η) as a metric for model evaluation and comparison. For the two models in Figure 2, with $\eta = .90$, the aforementioned AUCs are 97.8 and 75.4 for the models with and without baseline RNA, respectively. The corresponding ABCs are 17.3 and -5.1, respectively. Note that the ABC using the scoring system with baseline log(CD4) alone is negative, indicating that the overall performance of this scoring system is worse than a scoring system which groups the patients at random.

Now, if one considers the area under a weighted version of the curve, $(1-q)\tilde{AD}(q)$, this quantity consistently estimates

$$E\left(D(Z^0)\bar{H}(\bar{D}(Z^0))\right). \tag{3.2}$$

The expected value given in (3.2) directly measures the concordance of the subject-specific true treatment difference and the rank of the limiting score. This quantity may be easier to interpret heuristically than (3.1). Note that the performance of a scoring system only depends on the ranks of its scores. Moreover, the corresponding area between this curve and the straight line $y = (1-q)\tilde{AD}(0)$ is the covariance associated with the quantity given in (3.2) (see Appendix C for details). Also note that there are no existing procedures in the literature which can estimate such concordance measures at the patient level. Furthermore, if we could use the true treatment difference $D(Z)$ as the score, each of these concordance scores would attain its maximum value among all possible models derived from Z (see Appendix B for details).

When the dimension of the covariate vector Z is not small, it may not be appropriate to use the same data set to build a score via a complex variable selection algorithm and then use the same set to obtain $\tilde{AD}(\cdot)$ for model evaluation. Rather, one may randomly divide the data set into two independent pieces, the training and the evaluation sets, to avoid potential bias in assessing the adequacy of the model. When the data set is not large, an alternative approach is to use a random cross-validation procedure. Specifically, consider a class of models for the response Y and covariate vector Z. For each variable selection and estimation algorithm for this class of models, we randomly split the data set into two pieces, use the training set to obtain the scoring system $\hat{D}(Z)$, and construct the corresponding estimate $\hat{AD}(\cdot)$ using the evaluation set. We repeat this process M times.

Now, for the mth iteration, $m = 1, \cdots, M$, let $\hat{D}_m(Z)$, $\hat{AD}_m(c)$ and $H_m(c)$ be the corresponding aforementioned $\hat{D}(Z)$, $\hat{AD}(c)$ and $H(c)$, respectively. Let $\hat{D}_a(Z) = \frac{1}{M}\sum_{m=1}^{M}\hat{D}_m(Z)$, $\hat{AD}_a(c) = \frac{1}{M}\sum_{m=1}^{M}\hat{AD}_m(c)$, and $H_a(c) = \frac{1}{M}\sum_{m=1}^{M}H_m(c)$. Then $\tilde{AD}_a(q) = \hat{AD}_a(H_a^{-1}(q))$.

The comparisons among all the candidate models can be made via $\tilde{AD}_a(q)$. We then use the corresponding $\hat{AD}_a(c)$ of the best model to select the desirable subpopulation. Note that the score of a future subject with the covariate vector Z^0 is $\hat{D}_a(Z^0)$. This cross-training-evaluation averaging process is similar to bagging (Breiman, 1996).

We have conducted a large simulation study to examine the performance of the above cross-training-evaluation process. We find that under various practical settings, for each fitted model to create the scoring system, the empirical average of $\hat{AD}_a(\cdot)$ is almost identical to $AD(\cdot)$. Therefore, one can select the model using $\hat{AD}_a(\cdot)$. Moreover, the average score $\hat{D}_a(Z^0)$ to be utilized for selection of future study subjects gives us, for example, almost the same average treatment difference $E\left((Y^0_{(1)} - Y^0_{(0)})|\hat{D}_a(Z^0) \geq c\right)$ as $AD(c)$. More details of our numerical study results are given in the Remarks Section.

4. CREATING SCORING SYSTEM CANDIDATES

In this section, we discuss various models and variable selection procedures to build models for creating the scoring systems, for example, using the training data set for each iteration of the above cross-training-evaluation process. We first consider the case that (Y, G, Z) is completely observed. A general approach for modeling the subject-specific treatment difference parametrically is to model the mean for each treatment group:

$$\mu_k(Z) = g_k(\beta_k' h(Z)), \tag{4.1}$$

where $h(Z)$ is a known vector function of Z with the first component being 1, β_k is an unknown vector of parameters, g_k is a given link function, and $k = 0, 1$. To estimate β_k, one may minimize a loss function $L_k(\beta)$, which may be based on a likelihood or a residual sum of squares.

An alternative approach is to utilize a single model for both treatment groups:

$$E(Y|Z, G) = g(\beta' h(G, Z)), \tag{4.2}$$

where $h(G, Z)$ is a known vector function of (G, Z) with the first component being 1, β is an unknown vector of parameters, and g is a given link function. Note that $h(G, Z)$ may include G, Z, and $G \times Z$ interaction terms. In the presence of $G \times Z$ interaction terms, the results of variable selection procedures can change depending on how one codes the treatment indicator G. To this end, we code treatment group 0 and treatment group 1 using -1 and +1, respectively. Under this setting, $\hat{D}(Z) = g(\hat{\beta}' h(1, Z)) - g(\hat{\beta}' h(-1, Z))$. Again, one may obtain an estimator $\hat{\beta}$ for β by minimizing a loss function $L(\beta)$.

For Model (4.1) or (4.2), one may also use an estimation procedure for β with a built-in variable selection algorithm. For instance, for (4.2), let $\hat{\beta}_\lambda$ be a minimizer of

$$L(\beta) + \lambda \parallel \beta \parallel_d, \tag{4.3}$$

where $L(\beta)$ may be the negative log of the likelihood function for (4.2) or the residual sum of squares, and $\lambda > 0$ is the regularization parameter. Note that for the lasso procedure (Tibshirani, 1996), $d = 1$ and for ridge regression (Hoerl and Kennard, 1970), $d = 2$. One may select the regularization parameter $\hat{\lambda}$ based on the standard cross-validation procedure (Tibshirani, 1996). With the resulting $\hat{\beta}_{\hat{\lambda}}$, let $\hat{D}(Z)$ be the score.

Note that with a procedure using (4.3), it can be shown that, when the dimension of the covariate vector p is fixed and $\hat{\lambda} = o(n)$, $\hat{\beta}_{\hat{\lambda}}$ converges to a constant vector as $n \to \infty$ (Knight and Fu, 2000). This is an important property to guarantee that we will have a unique, well-defined limiting working model when repeating the algorithm with different training sets, as discussed in the previous section. Similarly, we may use the above variable selection algorithms with the model described in (4.1) separately for each treatment group. Similar to the L_d penalized estimator, the regression parameter estimator based on the standard stepwise variable selection procedure also has this stabilization property under more rigorous regularity conditions.

Now, consider the case that Y may not be observed completely due to censoring of the event time T. A common approach is to relate the event time to the covariates with a Cox proportional hazards model. For example, one may combine the data from two treatment groups and consider a working model:

$$\mathrm{pr}(T > t|Z, G) = g(\log \Lambda(t) + \beta' h(G, Z)), \tag{4.4}$$

where $g(x) = e^{-e^x}$, $h(G, Z)$ is a known vector function of (G, Z), $\Lambda(\cdot)$ is an unknown baseline cumulative hazard function, and β is an unknown vector of parameters. Again $h(G, Z)$ may include G, Z, and $G \times Z$ interaction terms. To estimate β, one may use the partial likelihood estimate. Here the loss function $L(\beta)$ is the negative log of the partial likelihood. An alternative is to utilize a corresponding (4.3) to obtain $\hat{\beta}_{\hat{\lambda}}$. Now, suppose $Y = I(T \geq t_0)$, where t_0 is a given time point. Then one may use

$$\hat{D}(Z) = g(\log \hat{\Lambda}(t_0) + \hat{\beta}'_{\hat{\lambda}} h(1, Z)) - g(\log \hat{\Lambda}(t_0) + \hat{\beta}'_{\hat{\lambda}} h(-1, Z)), \tag{4.5}$$

where

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s) e^{\hat{\beta}'_{\hat{\lambda}} h(G_j, Z_j)}}, \qquad (4.6).$$

with $N_i(t) = I(X_i \leq t)\Delta_i$ and $Y_i(t) = I(X_i \geq t)$, $i = 1, \ldots, n$.

If we are interested in the restricted mean event time, that is, $Y = \min(T, \tau_0)$, the resulting score from Model (4.4) is

$$\hat{D}(Z) = \int_0^{\tau_0} \left\{ g(\log \hat{\Lambda}(t) + \hat{\beta}'_{\hat{\lambda}} h(1, Z)) - g(\log \hat{\Lambda}(t) + \hat{\beta}'_{\hat{\lambda}} h(-1, Z)) \right\} dt. \tag{4.7}$$

Note that one may also fit a separate Cox model for each treatment group to create $\hat{D}(\cdot)$.

5. EXAMPLES

First, we illustrate our proposal using the data from the ACTG 320 HIV study described in the Introduction, using the nine baseline covariates listed in Table 1 of Hammer et al.

(1997). This set of covariates includes the baseline CD4 and RNA values. There are 838 patients who had complete information with respect to these 9 covariates. Again, we used week 24 CD4 value as the response variable Y, as in Section 2. Here, we consider two classes of models to construct various scoring systems. The first one, as in (4.1), uses an additive linear model for each treatment group with all nine of the covariates. The second one, as in (4.2), uses a single model with main covariate effects and interactions between the treatment indicator and other covariates. For each of the two classes of models, we used four variable selection procedures to build candidate scoring systems. For the first procedure, we used the full model with all the baseline covariates. For the second one, we used a stepwise variable selection based on Akaike information criterion (AIC) (Akaike, 1973). We then used lasso and ridge regression as the third and fourth variable selection procedures, respectively. The tuning parameters were selected by the standard cross-validation procedure built in the R package *glmnet*. For comparison, we also considered the two-variable model, discussed in Section 2, which uses only baseline CD4 and RNA.

Figure 3 summarizes the treatment difference curves $\tilde{AD}_a(\cdot)$ based on the averages over $M = 500$ replications of a cross-validation procedure, where each replication resulted from the random selection of 4/5 of the data as the training set. The results from these two classes of models are quite similar, except when using the lasso variable selection procedure. The model using only CD4 and RNA without variable selection performs well. On the other hand, the scoring systems using 9 covariates with the standard variable selection algorithms do not perform as well.

Now, if one wants to identify a subpopulation with an average CD4 count treatment difference of 100 cells/mm^3, then clearly the scoring system built with CD4 and RNA, which gives us the largest target subset of patients among all the candidate models, is the most favorable. In fact, using the two-variable model, 52% of the patients meet this criteria, while no more than 30% of the patient population is identified via any of the other candidate models. If this specific level of treatment difference is not of particular clinical interest, one may use the AUC and ABC discussed in Section 3 to compare the scoring systems. For example, with two separate models and $\eta = .90$, the ABC for the scoring system built using nine covariates with the lasso procedure is 9.1, compared with 17.3 for the simple model built using only CD4 and RNA.

As a second example, we considered a recent clinical trial "Prevention of Events with Angiotensin Converting Enzyme Inhibition" (PEACE) to study whether the ACE inhibitors (ACEi) are effective for reducing certain future cardiovascular-related events for patients with stable coronary artery disease and normal or slightly reduced left ventricular function (Braunwald et al., 2004). In this study, 4158 and 4132 patients were randomly assigned to the ACEi treatment and placebo arms, respectively. The median follow-up time was 4.8 years. One main endpoint for the study was the patient's survival time. By the end of the study, 334 and 299 deaths occurred in the control and treatment arms, respectively. Under a proportional hazards model, the estimated hazard ratio is 0.89 with a 0.95 confidence interval of (0.76, 1.04) and a p-value of 0.13. Based on the results of this study, it is not clear whether ACEi therapy would help the overall patient population with respect to mortality. However, with further analysis of the PEACE survival data, Solomon et al. (2006) reported that ACEi might significantly prolong survival for the subset of patients whose kidney functions at the study entry time were not normal (for example, those with estimated glomerular filtration rate, eGFR, < 60). This finding could be quite useful in practice. On the other hand, such a subgroup analysis has to be executed properly and the results of such analysis have to be interpreted cautiously (Rothwell, 2005; Pfeffer and Jarcho, 2006; Wang et al., 2007).

For this example, we considered the time-to-event endpoint, T, the time to all-cause mortality. To build a candidate scoring system, we first used the 7 covariates previously identified as statistically and clinically important predictors of the overall mortality in the literature (Solomon et al., 2006). These covariates are eGFR, age, gender, left ventricular ejection fraction (lveejf), history of hypertension, diabetes, and history of myocardial infarction. For comparison, we also used two scoring systems built using eGFR alone and lveejf alone, which are two conventional predictive markers for cardiovascular diseases. In addition, we considered the scoring systems built with various variable selection procedures using the baseline covariates listed in Table 2 of Braunwald et al. (2004). However, we did not use three of the variables listed: race, country, and serum creatinine, which were not available in our database from the US National Institutes of Health. Moreover, we omitted four binary variables due to lack of variability (i.e., over 95% of patients exhibited the same covariate value). These excluded variable are: use of Digitalis, use of antiarrhythmic agent, use of anticoagulant, and use of insulin. On the other hand, an extra variable eGFR, which is a function of age, gender, race, and serum creatinine, was available in our database. To this end, we considered the remaining 20 variables from Table 2 of Braunwald et al. (2004) in addition to eGFR, resulting in a total of 21 covariates. In our analysis, we included all patients ($n = 7460$) who had complete information concerning these 21 covariates. To estimate the score for the treatment differences, we considered two classes of models: a separate Cox model for each of the two treatment groups, and a single Cox model which includes treatment-covariate interaction terms. For each of the two classes of models, we used the same four variable selection procedures as in the previous example to build candidate scoring systems.

First, suppose that one is interested in survival probability at month 72. We let $Y = I(T \geq 72)$. Figure 4 summarizes the treatment difference curves for various scoring systems based on 500 random cross-validations with 4/5 of the data as the training set. The treatment difference curve with the 7 clinically meaningful covariates and the one with eGFR alone are similar. Both perform uniformly better than any of the scoring systems which use all 21 covariates. When using two separate models, as shown in the left panel of Figure 4, the performance of the scoring systems constructed via variable selection procedures appears similar to the full model. Using a single interaction model (right panel), the stepwise and lasso variable selection procedures appear inferior to the one with all 21 covariates. It is interesting to note that the scoring system based on lveejf alone performs quite poorly, indicating that this conventional marker for cardiovascular diseases by itself is not helpful in identifying patients who would benefit from ACEi. To further quantify the relative performance among the candidate scoring systems, one may use the AUC and ABC discussed in Section 3. For example, with two separate models and $\eta = .90$, the ABC for the scoring system built with 7 covariates is 0.015, which is the largest among all candidates. The estimated ratio of correlations between the true treatment difference $D(Z^0)$ and $\log\{[1 - \bar{H}(\bar{D}(Z^0))]^{-1}\}$ using this scoring system is 1.21 relative to that using eGFR alone, 1.65 relative to the one using all 21 covariates, and 4.11 relative to that using lveejf alone.

Next, suppose that one is interested in the restricted mean event time up to month 72. To this end, we let $Y = \min(T, 72)$. Figure 5 presents the results based on 500 random cross-validations with 4/5 of the data as the training set. The scoring system built with the 7 covariates appears to outperform the others. Again it appears that the scoring systems created using the variable selection procedures with 21 covariates perform similarly or inferior to the one with the full model, and the system based on lveejf only performs poorly. It is interesting to note that the model with eGFR alone does not perform particularly well for this endpoint.

Based on the partial AUC and ABC, the scoring system using two separate models with 7 covariates is the best among the candidate models for the survival probability at month 72. This model also gives the best scoring system among the candidate models for the restricted mean event time up to month 72. Figure 6 provides the estimated average treatment differences $\hat{AD}_a(c)$ over a range of values c for both endpoints. From this figure, one can easily identify the subgroup of patients with any desired level of treatment benefit. For example, if we desire a 72-month survival rate benefit of 0.05, since $\hat{AD}(0.038) = 0.05$, we can identify the subset of patients with Z^0 such that $\hat{D}(Z^0) \geq 0.038$. If we desire a treatment benefit of 1.5 months for the restricted mean event time up to month 72, the corresponding subset could consist of patients Z^0 such that $\hat{D}(Z^0) \geq 2.23$.

6. REMARKS

Note that the typical subgroup analysis strategy, which tries to identify a target population for future study by dichotomizing one or more baseline variables, may not be efficient, especially when the dimension of Z is large. That is, the resulting population selected by this strategy can be quite small, which is not practically useful. Our proposal selects the largest population whose subjects would have a desired overall treatment benefit, among all candidate scoring systems.

We conducted an extensive numerical study to examine the performance of the new proposal under various practical settings. We find that the empirical average of the estimates $\hat{AD}_a(c)$ via the random cross-validation procedure is practically identical to its theoretical counterpart $AD(c)$ in (2.1) when c is not very large (the upper tail of $\hat{AD}_a(\cdot)$ may not be stable). On the other hand, if we use the entire data set to fit a model for creating a scoring system, and use the same data set to estimate $AD(c)$, the resulting estimator $\hat{AD}(\cdot)$ can be expected to be substantially overly optimistic. For example, we mimicked the HIV example to generate the data for our numerical study. Specifically, we assumed a single linear model with response Y being the week 24 CD4 count and independent variables being the treatment indicator, the nine baseline covariates discussed in Section 5, and the treatment-covariate interactions. The error of the model was assumed to be normal with mean zero. We fitted the HIV study data using this model and then used this model to generate responses. To simulate a data set with sample size n, for each subject, first we randomly chose a covariate vector and the treatment indicator from the original study database with replacement. We then generated a week 24 CD4 count using the above "true" model. For each simulated data set, we fitted two separate linear models, one for the control and one for the treatment group, using the above 9 covariates additively. With the resulting scores $\hat{D}(Z)$, we estimated the mean value of the treatment difference $Y^0_{(1)} - Y^0_{(0)}$ given $\hat{D}(Z) > c$, with 10,000 fresh independent observations (Y, G, Z). We replicated this process 100 times and used the empirical average to approximate $AD(c)$. The resulting curve (solid) is given in Figure 7 with $n = 838$. Now, to obtain an empirical average of $\hat{AD}_a(c)$, we used the above 100 simulated data sets with sample size n. The random cross-validation procedures were repeated 100 times for each simulated data set. The dashed curve is the resulting empirical average of $\hat{AD}_a(c)$ with a 4:1 ratio of training and evaluation samples. The dotted curve is the corresponding empirical average of $\hat{AD}(\cdot)$, where the same data set is used for both training and evaluation. Note that the dotted curve is markedly higher than the solid one, indicating that the procedure using the entire data set for model building and evaluation can be quite misleading. From our extensive numerical study, we find that the estimation procedure for $AD(\cdot)$ performs well with a random K-fold cross validation when $5 \leq K \leq 10$ (that is, using $(K-1)/K$ as training and $1/K$ as evaluation repeatedly).

It is important to note that it is difficult, if not impossible, to make further inferences, for example, constructing confidence intervals or bands, about the average treatment difference curve $AD(\cdot)$ using only a single data set. This is due to the fact that we perform a large number of model building and selection processes to identify the best scoring system. The sampling variation for the final estimator $\hat{AD}_a(\cdot)$ cannot be derived based on the conventional fixed model fitting procedure. If there is an independent data set generated from a similar population, the techniques for analyzing standard empirical processes may be utilized for constructing the interval estimates by treating the scores as being fixed (Song and Pepe, 2004; Song and Zhou, 2011).

From a risk-benefit perspective for evaluating the new treatment, one may collect toxicity information and construct corresponding treatment contrast measures using the same efficacy score $\hat{D}(Z)$. The resulting two sets of curves can be quite useful for selecting a proper target population. For comparing multiple treatment arms with a control, we may construct pair-wise treatment-control difference curves $\tilde{AD}_a(\cdot)$. It follows from our proposal that the treatment with the highest treatment difference curve or a function thereof may be selected to be the candidate for the future studies.

References

Akaike, H. (**1973**), "Information theory and an extension of the maximum likelihood principle," in *Second international symposium on information theory* (Springer Verlag), vol. 1, pp. 267–281.

Andersen, P., Hansen, M., and Klein, J. (**2004**), "Regression analysis of restricted mean survival time based on pseudo-observations," Lifetime Data Analysis **10**(4), 335–350.

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

Braunwald, E., Domanski, M. J., Fowler, S. E., and et al., The PEACE Trial Investigators (**2004**), "Angiotensin-converting-enzyme inhibition in stable coronary artery disease," The New England Journal of Medicine **351**, 2058–2068.

Breiman, L. (**1996**), "Bagging predictors," Machine learning **24**(2), 123–140.

Cai, T., Tian, L., Wong, P., and Wei, L. (**2011**), "Analysis of randomized comparative clinical trial data for personalized treatment selections," Biostatistics **12**(2), 270.

Cox, D. R. (**1972**), "Regression models and life-tables (with discussion)," Journal of the Royal Statistical Society, Series B **34**, 187–220.

Hammer, S., Squires, K., Hughes, M., Grimes, J., Demeter, L., Currier, J., Eron, J., Feinberg, J., Balfour, H., Deyton, L., *et al.* (**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," New England Journal of Medicine-Unbound Volume **337**(11), 725–733.

Hoerl, A. and Kennard, R. (**1970**), "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics **12**(1), 55–67.

Irwin, J. (**1949**), "The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice," Journal of Hygiene **47**(02), 188–189.

Janes, H., Pepe, M., Bossuyt, P., and Barlow, W. (**2011**), "Measuring the Performance of Markers for Guiding Treatment Decisions," Annals of internal medicine **154**(4), 253.

Kalbfleisch, J. D. and Prentice, R. L. (**1981**), "Estimation of the average hazard ratio," Biometrika **68**, 105–112.

Kalbfleisch, J. D. and Prentice, R. L. (**2002**), *The Statistical Analysis of Failure Time Data* (New York: JohnWiley & Sons).

Knight, K. and Fu, W. (**2000**), "Asymptotics for lasso-type estimators," The Annals of Statistics **28**(5), 1356–1378.

Lin, D. Y. and Wei, L. J. (**1989**), "The robust inference for the Cox proportional hazards model," Journal of American Statistical Association **84**, 1074–1078.

Moskowitz, C. and Pepe, M. (**2004**), "Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes," Biostatistics **5**(1), 113.

Pepe, M. S. (**2003**), *The statistical evaluation of medical tests for classification and prediction* (Oxford University Press).

Pfeffer, M. and Jarcho, J. (**2006**), "The Charisma of Subgroups and the Subgroups of CHARISMA," New England Journal of Medicine **354**(16), 1744.

Pollard, D. (**1990**), *Empirical Processes: Theory and Applications, Regional Conference Series in Probability and Statistics 2* (Institute of Mathematical Statistics, Hayward, CA).

Rothwell, P. (**2005**), "External validity of randomised controlled trials: "to whom do the results of this trial apply?"," The Lancet **365**(9453), 82–93.

Solomon, S. D., M., R. M., Jablonski, K. A., and et al., for the Prevention of Events with ACE inhibition (PEACE) Investigators (**2006**), "Renal Function and Effectiveness of Angiotensin-Converting Enzyme Inhibitor Therapy in Patients With Chronic Stable Coronary Disease in the Prevention of Events with ACE inhibition (PEACE) Trial," Circulation **114**, 26–31.

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

Song, X. and Zhou, X. (**2011**), "Evaluating Markers for Treatment Selection Based on Survival Time," UW Biostatistics Working Paper Series , 375.

Tibshirani, R. (**1996**), "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B (Methodological) **58**(1), 267–288.

Wang, R., Lagakos, S., Ware, J., Hunter, D., and Drazen, J. (**2007**), "Statistics in Medicine–Reporting of Subgroup Analyses in Clinical Trials," New England Journal of Medicine **357**(21), 2189.

Xu, R. and O'Quigley, J. (**2000**), "Estimating Average Regression Effect under Non-Proportional Hazards," Biostatistics **1**, 423–439.

APPENDIX A: CONSISTENCY OF $\hat{AD}(\cdot)$

In this appendix, we outline the proof for (2.5) and (2.6). We assume that the covariate vector $Z^0 \in \Omega$ is bounded. In addition, we assume that $\hat{D}(Z^0)$ converges in probability to a finite constant $\bar{D}(Z^0)$ uniformly in $Z^0 \in \Omega$, as $n \to \infty$. This assumption holds for all the scoring systems considered in Section 4, because of the convergence of the regression parameter estimator $\hat{\beta}$ regardless of whether the working model is correctly specified (Cai et al., 2011). Note that $\bar{D}(Z^0)$ could be different from $D(Z^0)$ when the working model is misspecified. Without loss of generality, we assume that at least one of the covariates is continuous, but similar arguments can be used to justify the discrete case. Furthermore, we assume that $D(Z^0)$ and $\bar{D}(Z^0)$ are bounded, and the probability density function of $\bar{D}(Z^0)$ is bounded.

We first prove (2.5). To this end, let $\epsilon = \sup_{Z^0 \in \Omega} |\hat{D}(Z^0) - \bar{D}(Z^0)| = o_p(1)$ by assumption. It is easy to show that

$$\left| I\{\hat{D}(Z^0) \geq c\} - I\{\bar{D}(Z^0) \geq c\} \right| \leq I\left\{\bar{D}(Z^0) \in [c-\epsilon, c+\epsilon)\right\}.$$

Because the probability density function of $\bar{D}(Z^0)$ is bounded, and $\epsilon = o_p(1)$, it follows that

$$\sup_{c \in (-\infty, c_0)} \left| \mathrm{pr}(\hat{D}(Z^0) \geq c) - \mathrm{pr}(\bar{D}(Z^0) \geq c) \right| \leq \sup_{c \in (-\infty, c_0)} \mathrm{pr}\left(\bar{D}(Z^0) \in [c-\epsilon, c+\epsilon)\right) = o(1). \tag{A.1}$$

Similarly, suppose $|D(Z^0)|$ is bounded by B_0, we have

$$\begin{aligned} &\sup_{c \in (-\infty, c_0)} \left| E\left(D(Z^0) I\{\hat{D}(Z^0) \geq c\}\right) - E\left(D(Z^0) I\{\bar{D}(Z^0) \geq c\}\right) \right| \\ &\qquad \leq B_0 \sup_{c \in (-\infty, c_0)} \mathrm{pr}\left(\bar{D}(Z^0) \in [c-\epsilon, c+\epsilon)\right) = o(1) \end{aligned} \tag{A.2}$$

Combining (A.1) and (A.2), and the fact that $AD(c) = E\left(D(Z^0) I\{\hat{D}(Z^0) \geq c\}\right) / \mathrm{pr}(\hat{D}(Z^0) \geq c)$ and $\bar{AD}(c) = E\left(D(Z^0) I\{\bar{D}(Z^0) \geq c\}\right) / \mathrm{pr}(\bar{D}(Z^0) \geq c)$, it follows that (2.5) is true.

Next, we outline the proof for (2.6) separately for the cases when the response variable Y is completely observed and not completely observed. We first consider the case when Y is completely observed. For simplicity, we assume Y is bounded. By the uniform law of large numbers (Pollard, 1990), we have

$$\sup_{c\in(-\infty,c_0)}\left|\frac{\sum_i^n Y_i I\{\bar{D}(Z_i)\geq c, G_i=k\}}{\sum_i^n I\{\bar{D}(Z_i)\geq c, G_i=k\}} - E\left(Y^0_{(k)}|\bar{D}(Z^0)\geq c\right)\right| = o_p(1),$$

for $k=1,2$. Thus it suffices to show that

$$\sup_{c\in(-\infty,c_0)}\left|\frac{\sum_i^n Y_i I\{\hat{D}(Z_i)\geq c, G_i=k\}}{\sum_i^n I\{\hat{D}(Z_i)\geq c, G_i=k\}} - \frac{\sum_i^n Y_i I\{\bar{D}(Z_i)\geq c, G_i=k\}}{\sum_i^n I\{\bar{D}(Z_i)\geq c, G_i=k\}}\right| = o_p(1), \qquad (A.3)$$

for $k=1,2$. It is easy to show that

$$\left|I\{\hat{D}(Z_i)\geq c, G_i=k\} - I\{\bar{D}(Z_i)\geq c, G_i=k\}\right| \leq I\left\{\bar{D}(Z_i)\in[c-\epsilon, c+\epsilon), G_i=k\right\}. \qquad (A.4)$$

Because the probability density function of $\bar{D}(Z^0)$ is bounded, and $\epsilon = o_p(1)$, by the uniform law of large numbers (Pollard, 1990), we have

$$\sup_{c\in(-\infty,c_0)} \frac{1}{n\pi_k}\sum_{i=1}^n I\left\{\bar{D}(Z_i)\in[c-\epsilon, c+\epsilon), G_i=k\right\} = o_p(1). \qquad (A.5)$$

Combining (A.4) and (A.5), it follows that

$$\sup_{c\in(-\infty,c_0)}\left|\frac{1}{n\pi_k}\sum_i^n I\{\hat{D}(Z_i)\geq c, G_i=k\} - \frac{1}{n\pi_k}\sum_i^n I\{\bar{D}(Z_i)\geq c, G_i=k\}\right| = o_p(1).$$

Suppose $|Y_i|\leq B_1$. Similar to (A.4), we have

$$\left|Y_i I\{\hat{D}(Z_i)\geq c, G_i=k\} - Y_i I\{\bar{D}(Z_i)\geq c, G_i=k\}\right| \leq B_1 I\left\{\bar{D}(Z_i)\in[c-\epsilon, c+\epsilon), G_i=k\right\}.$$

It follows by the same arguments as above that

$$\sup_{c\in(-\infty,c_0)}\left|\frac{1}{n\pi_k}\sum_i^n Y_i I\{\hat{D}(Z_i)\geq c, G_i=k\} - \frac{1}{n\pi_k}\sum_i^n Y_i I\{\bar{D}(Z_i)\geq c, G_i=k\}\right| = o_p(1).$$

Thus (A.3) is true. This proves (2.6) when Y is completely observed.

Now, we prove (2.6) for the case when Y is not completely observed. Here we consider the response for $Y = I(T \geq t_0)$. The proof for $Y = \min(T, \tau_0)$ is similar. Let $\bar{N}_{i,c}^{(k)}(t) = I(X_i \leq t, \bar{D}(Z_i) \geq c, G_i = k)\Delta_i$, and $\bar{Y}_{i,c}^{(k)}(t) = I(X_i \geq t, \bar{D}(Z_i) \geq c, G_i = k)$, $k = 0, 1$; $i = 1, \ldots, n$. Firstly, following the equivalence between the Kaplan-Meier estimator and the Nelson-Aalen estimator, it is not difficult to see that

$$\prod_{t=0}^{t_0} \left\{ 1 - \frac{\sum_{i=1}^n dN_{i,c}^{(k)}(t)}{\sum_{i=1}^n Y_{i,c}^{(k)}(t)} \right\} = \exp \left\{ -\frac{1}{n} \sum_{i=1}^n \int_0^{t_0} \frac{dN_{i,c}^{(k)}(t)}{L_c^{(k)}(t)} \right\} + o_p(1),$$

and

$$\prod_{t=0}^{t_0} \left\{ 1 - \frac{\sum_{i=1}^n d\bar{N}_{i,c}^{(k)}(t)}{\sum_{i=1}^n \bar{Y}_{i,c}^{(k)}(t)} \right\} = \exp \left\{ -\frac{1}{n} \sum_{i=1}^n \int_0^{t_0} \frac{d\bar{N}_{i,c}^{(k)}(t)}{\bar{L}_c^{(k)}(t)} \right\} + o_p(1),$$

uniformly for $c \in (-\infty, c_0)$, where $L_c^{(k)}(t) = \mathrm{pr}(X_i \geq t, \hat{D}(Z_i) \geq c, G_i = k)$ and $\bar{L}_c^{(k)}(t) = \mathrm{pr}(X_i \geq t, \bar{D}(Z_i) \geq c, G_i = k)$. Note that the above two exponential terms are the sum of n iid terms. The rest of proof is similar to the completely observed case.

APPENDIX B: OPTIMALITY OF THE TRUE SUBJECT-SPECIFIC TREATMENT DIFFERENCE AS A SCORING SYSTEM

It can be shown that $H(\hat{D}(Z^0))$ converges in probability to $\bar{H}(\bar{D}(Z^0))$ uniformly in $Z^0 \in \Omega$, as $n \to \infty$. Thus it follows by the same arguments of Appendix A that $\tilde{AD}(q)$ defined in Section 3 is a consistent estimator for $E(Y_{(1)}^0 - Y_{(0)}^0 | \bar{H}(\bar{D}(Z^0)) \geq q)$, which can be rewritten as

$$\frac{E\left(\left(Y_{(1)}^0 - Y_{(0)}^0\right) I\{\bar{H}(\bar{D}(Z^0)) \geq q\}\right)}{\mathrm{pr}\{\bar{H}(\bar{D}(Z^0)) \geq q\}} = \frac{\int I\{\bar{H}(\bar{D}(z)) \geq q\} D(z) f(z) dz}{1-q}, \qquad (B.1)$$

where $f(z)$ is the density function of Z^0.

Now we show that the limit of $\tilde{AD}(\cdot)$ is pointwise maximized when the true subject-specific treatment difference $D(Z^0)$ is used as the scoring system. From (B.1), the limit of

$\tilde{AD}(q)$ with the scoring system $\hat{D}(Z^0)$ is $\int I\{\bar{H}(\bar{D}(z)) \geq q\} D(z)f(z)dz/(1-q)$. Similarly, the limit of $\tilde{AD}(q)$ with the scoring system $D(Z^0)$ is $\int I\{H_0(D(z)) \geq q\} D(z)f(z)dz/(1-q)$, where $H_0(\cdot)$ is the cumulative distribution function of $D(Z^0)$. It follows that

$$
\begin{aligned}
& \int I\{H_0(D(z)) \geq q\} D(z)f(z)dz - \int I\{\bar{H}(\bar{D}(z)) \geq q\} D(z)f(z)dz \\
= & \int \left[I\left\{D(z) \geq H_0^{-1}(q)\right\} - I\left\{\bar{D}(z) \geq \bar{H}^{-1}(q)\right\}\right] D(z)f(z)dz \\
= & \int \left[I\left\{D(z) \geq H_0^{-1}(q), \bar{D}(z) < \bar{H}^{-1}(q)\right\}\right] D(z)f(z)dz - \\
& \int \left[I\left\{D(z) < H_0^{-1}(q), \bar{D}(z) \geq \bar{H}^{-1}(q)\right\}\right] D(z)f(z)dz \\
\geq & \int \left[I\left\{D(z) \geq H_0^{-1}(q), \bar{D}(z) < \bar{H}^{-1}(q)\right\}\right] H_0^{-1}(q)f(z)dz - \\
& \int \left[I\left\{D(z) < H_0^{-1}(q), \bar{D}(z) \geq \bar{H}^{-1}(q)\right\}\right] H_0^{-1}(q)f(z)dz \\
= & H_0^{-1}(q) \int \left[I\left\{D(z) \geq H_0^{-1}(q)\right\} - I\left\{\bar{D}(z) \geq \bar{H}^{-1}(q)\right\}\right] f(z)dz \\
= & 0.
\end{aligned}
$$

This proves that for each $q \in (0,1)$, the limit of $\tilde{AD}(q)$ is maximized when the true subject-specific treatment difference $D(Z^0)$ is used as the scoring system. Thus so are the limits of AUC, partial AUC, and weighted AUC of $\tilde{AD}(\cdot)$, and the corresponding ABCs.

APPENDIX C: AUC AND ABC OF WEIGHTED $\tilde{AD}(\cdot)$

Suppose that $w_0(q), q \in [0,1]$ is a non-negative weight function, in this section we will justify that

$$\int_0^\eta w_0(q)\tilde{AD}(q)dq \xrightarrow{p} E\left\{D(Z^0)\psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right\}. \qquad (C.1)$$

and

$$\int_0^\eta w_0(q)\left\{\tilde{AD}(q) - \tilde{AD}(0)\right\} dq \xrightarrow{p} \mathrm{Cov}\left(D(Z^0), \psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right), \qquad (C.2)$$

where

$$\psi_{w_0}(q) = \int_0^x \frac{w_0(q)dq}{1-q}$$

and $0 < \eta < 1$ is a fixed constant.

It follows from (B.1) that

$$\begin{aligned}
& \int_0^\eta w_0(q)\tilde{AD}(q)dq \\
\xrightarrow{p} & \int_0^\eta w_0(q)\left[\frac{\int I\{\bar{H}(\bar{D}(z)) \geq q\}D(z)f(z)dz}{1-q}\right]dq \\
= & \int \left[\int_0^{\min\{\eta,\bar{H}(\bar{D}(z))\}} \frac{w_0(q)}{1-q}dq\right] D(z)f(z)dz \\
= & \int \psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(z))\}\right) D(z)f(z)dz \\
= & E\left\{D(Z^0)\psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right\}
\end{aligned}$$

Therefore (C.1) is true. To justify (C.2), we first note that $\tilde{AD}(0)$ is consistent for $E(D(Z^0)) = E(Y_{(1)}) - E(Y_{(0)})$, which is simply the overall treatment difference. In addition,

$$\begin{aligned}
& E\left\{\psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right\} \\
= & \int_0^1 \psi_{w_0(q)}\left\{\min(\eta, u)\right\} du \\
= & \int_0^1 \int_0^{\min(\eta,u)} \frac{w_0(q)}{1-q}dqdu \\
= & \int_0^\eta w_0(q)dq,
\end{aligned}$$

where we used the fact that $\bar{H}\{\bar{D}(Z^0)\}$ follows uniform distribution. Coupled with (C.1), it follows that

$$\begin{aligned}
& \int_0^\eta w_0(q)\left\{\tilde{AD}(q) - \tilde{AD}(0)\right\}dq \\
\xrightarrow{p} & E\left\{D(Z^0)\psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right\} - E\{D(Z^0)\}E\left\{\psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right\} \\
= & \operatorname{Cov}\left(D(Z^0), \psi_{w_0}\left(\min\{\eta, \bar{H}(\bar{D}(Z^0))\}\right)\right),
\end{aligned}$$

which justifies (C.2).

When $w_0(q) = 1$, we have $\psi_{w_0}(q) = -\log\{(1-q)\}$, $\psi_{w_0}(\bar{H}(\bar{D}(Z^0)))$ follows unit exponential distribution, and

$$\int_0^1 \left\{\tilde{AD}(q) - \tilde{AD}(0)\right\} dq \xrightarrow{p} \sigma_0 \mathrm{Cor}\left(D(Z^0), \psi_{w_0}\left(\bar{H}(\bar{D}(Z^0))\right)\right),$$

where σ_0 is the standard deviation of $D(Z^0)$, an unknown constant which does not depend on any specific scoring system. If $w_0(q)$ is set as $1-q$, then $\psi_{w_0}(q) = q$ and

$$\int_0^1 (1-q)\left\{\tilde{AD}(q) - \tilde{AD}(0)\right\} dq \xrightarrow{p} \frac{\sigma_0}{2\sqrt{3}} \mathrm{Cor}\left(D(Z^0), \bar{H}\{\bar{D}(Z^0)\}\right),$$

where we use the fact that $\mathrm{Var}\{\bar{H}(\bar{D}(Z^0))\} = 1/12$ because $\bar{H}(\bar{D}(Z^0))$ follows standard uniform distribution.

Table 1: Estimated (Est) regression coefficients, their standard errors (SE) and p-values by fitting two separate linear regression models to the ACTG 320 data with week 24 CD4 as the response and log(CD4) and $\log_{10}$(RNA) as the baseline covariates

Covariates	Two-drug			Three-drug		
	Est	SE	p-value	Est	SE	p-value
Intercept	-17.04	24.13	0.48	-137.66	40.91	<0.01
log(CD4)	43.05	2.31	<0.01	55.62	3.83	<0.01
$\log_{10}$(RNA)	-9.98	4.05	0.01	19.16	6.85	0.01

Table 2: Estimated (Est) regression coefficients, their standard errors (SE) and p-values by fitting two separate linear regression models to the ACTG 320 data with week 24 CD4 as the response and log(CD4) alone as the baseline covariates

Covariates	Two-drug			Three-drug		
	Est	SE	p-value	Est	SE	p-value
Intercept	-72.21	9.08	<0.01	-31.64	15.57	0.04
log(CD4)	44.52	2.24	<0.01	52.79	3.72	<0.01

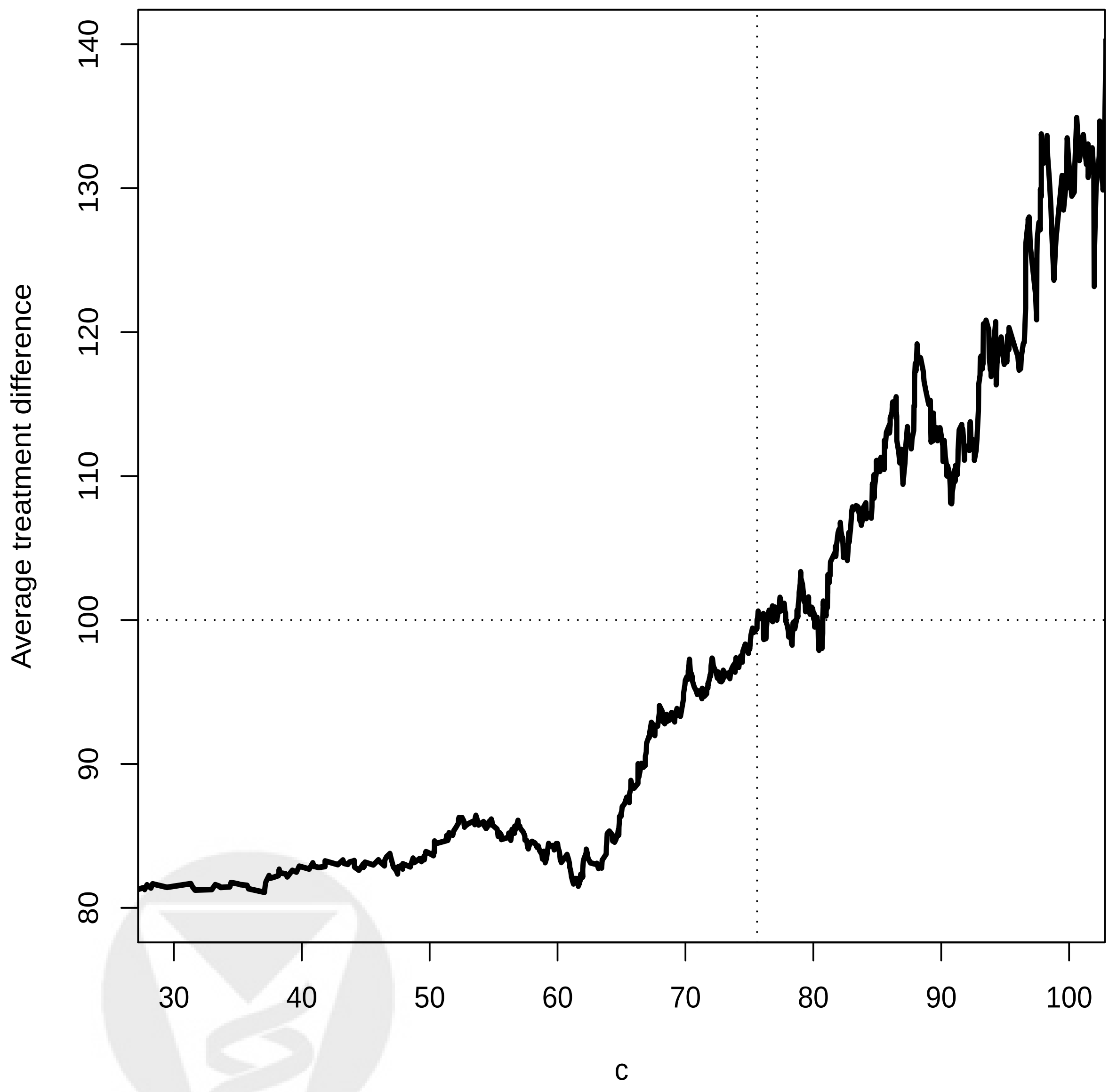

Figure 1: Estimated average treatment difference for patients with $\hat{D}(Z) \geq c$ using the scoring system built with two baseline covariates, log(CD4) and $\log_{10}$(RNA), for the ACTG 320 data

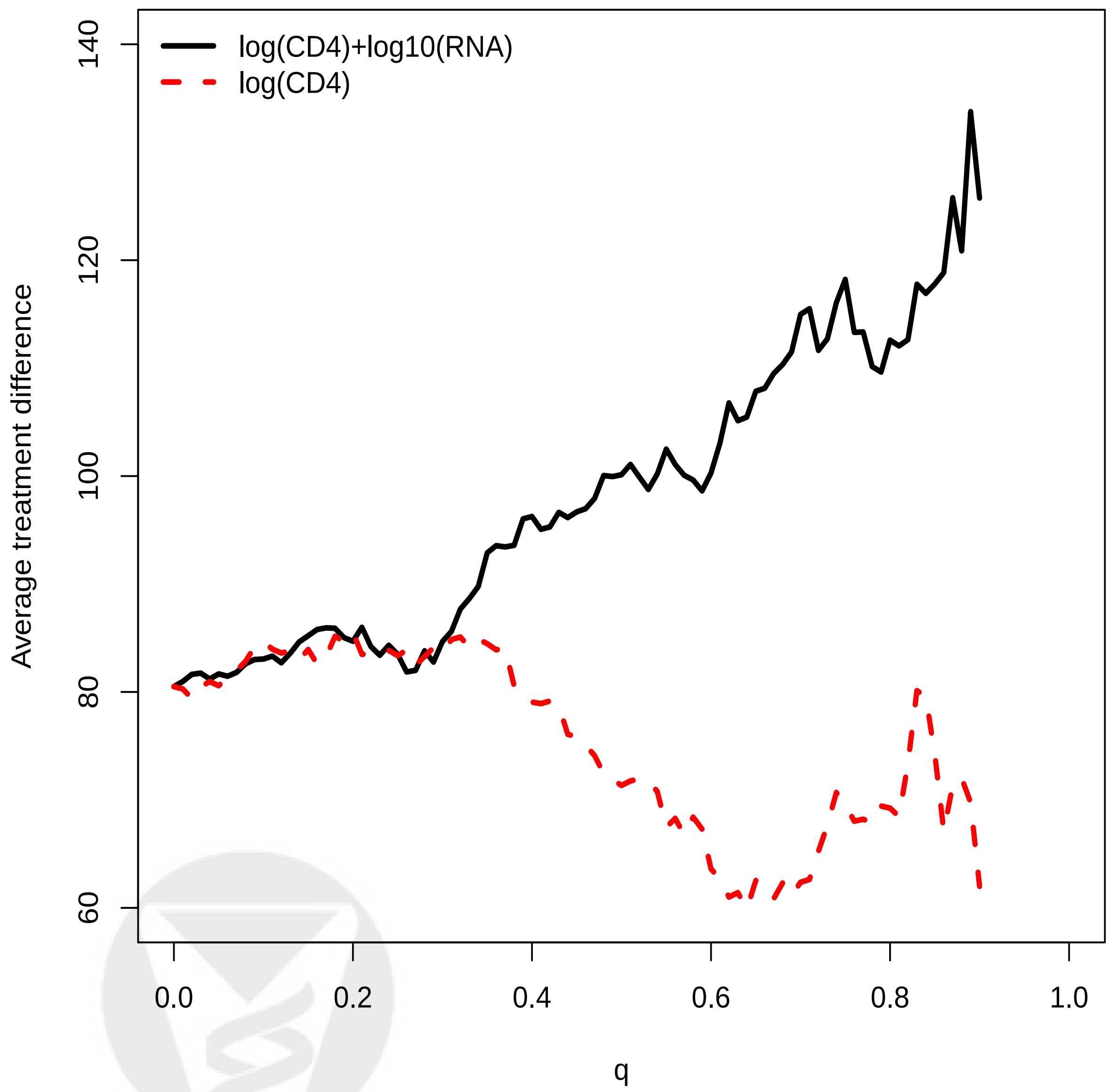

Figure 2: Comparing the two estimated average treatment differences for patients with largest $100(1-q)\%$ scores using the systems built with and without $\log_{10}$(RNA) for the ACTG 320 data

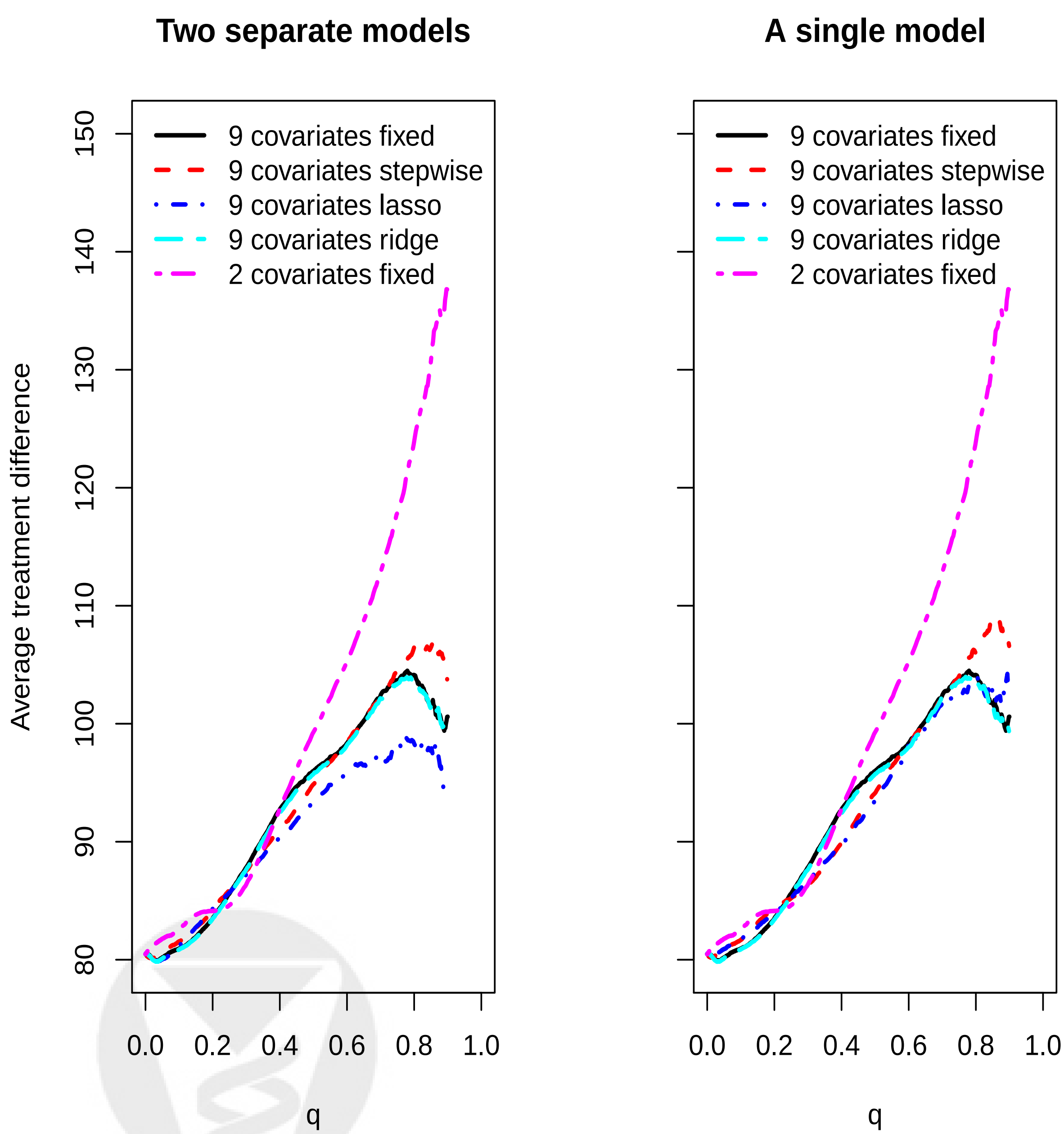

Figure 3: Comparing the estimated average treatment difference curves using various scoring systems based on 500 replicates of cross-validation for the ACTG 320 data (left panel: two separate models; right panel: a single interaction model)

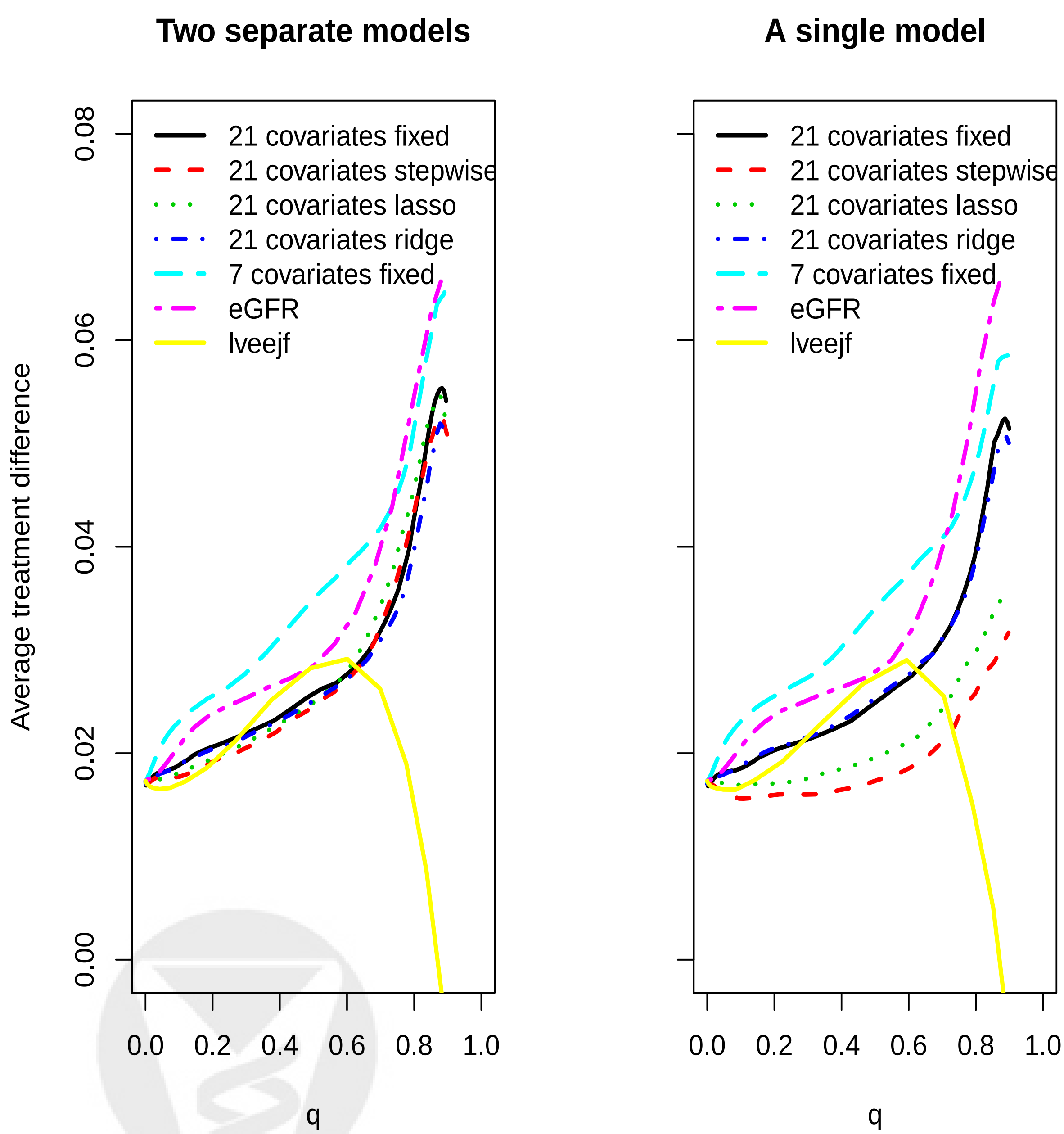

Figure 4: Comparing the estimated average treatment difference curves using different scoring systems with respect to 72-month survival rate, based on 500 replicates of cross-validation for the PEACE data (left panel: two separate models; right panel: a single interaction model)

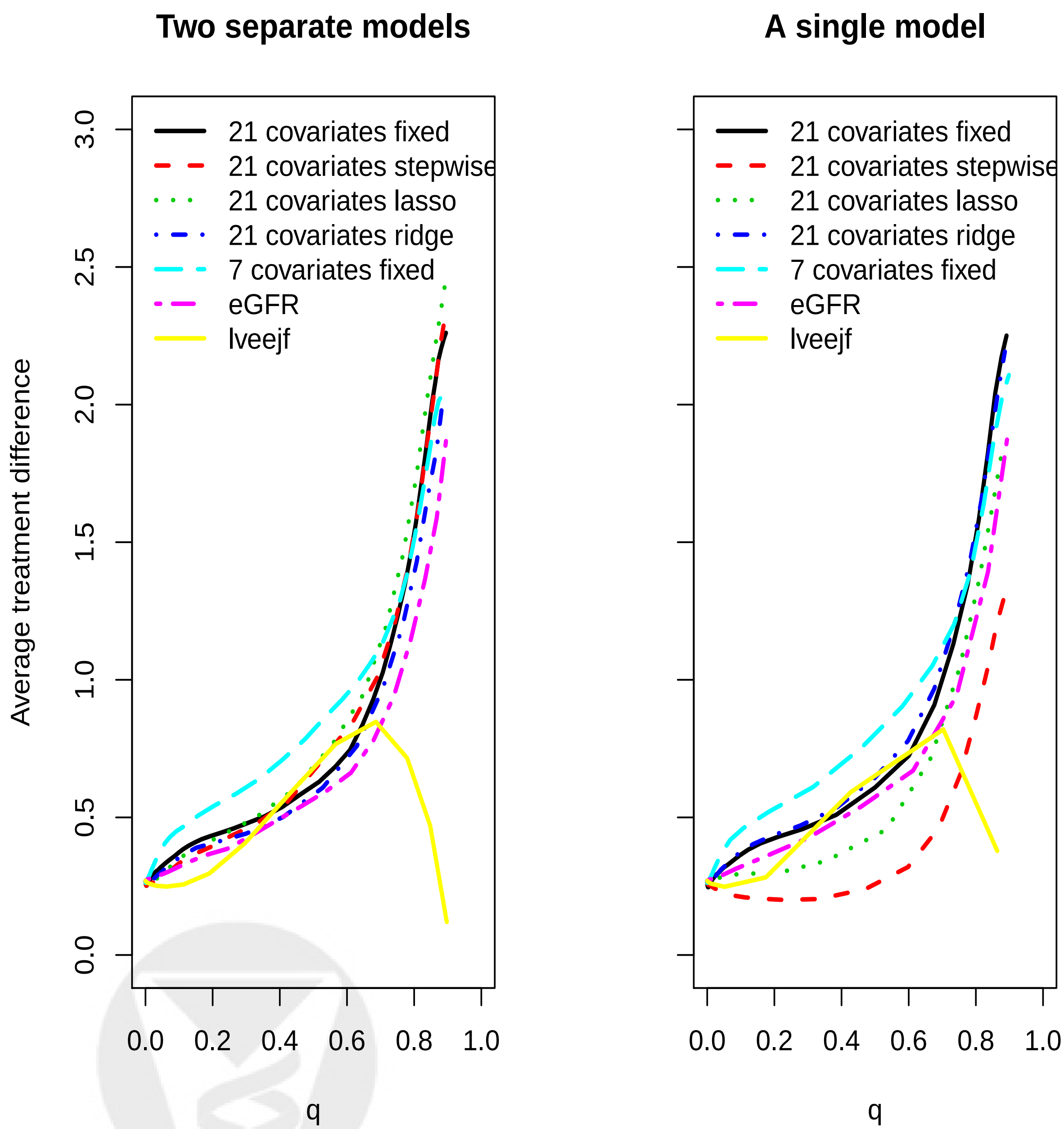

Figure 5: Comparing the estimated average treatment difference curves using different scoring systems with respect to restricted mean survival time up to 72 months, based on 500 replicates of cross-validation for the PEACE data (left panel: two separate models; right panel: a single interaction model)

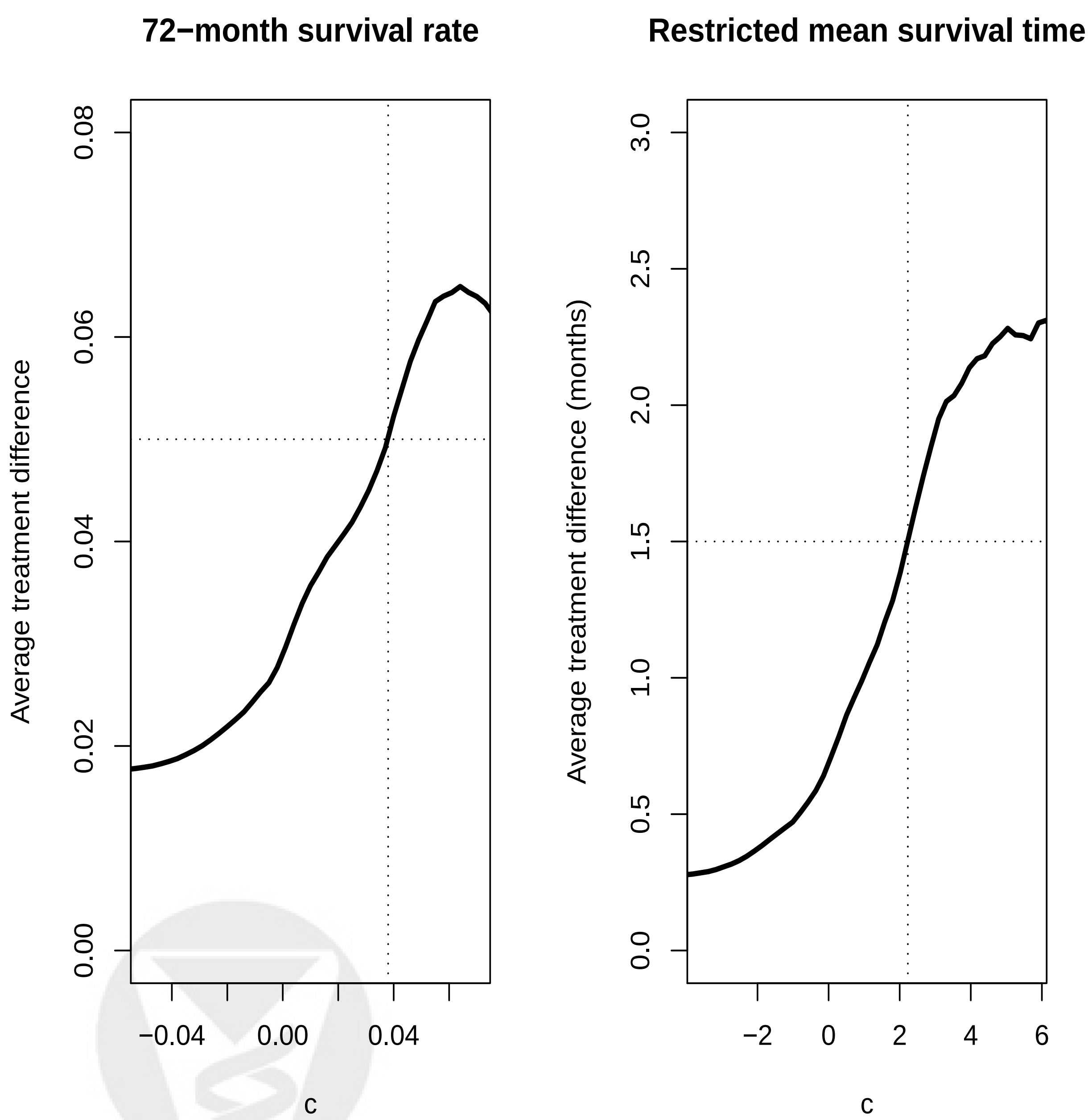

Figure 6: Estimated average treatment difference for patients with $\hat{D}(Z) \geq c$ using the scoring system built with two separate models and 7 covariates for the PEACE data (left panel: 72-month survival rate; right panel: restricted mean survival time up to 72 months)

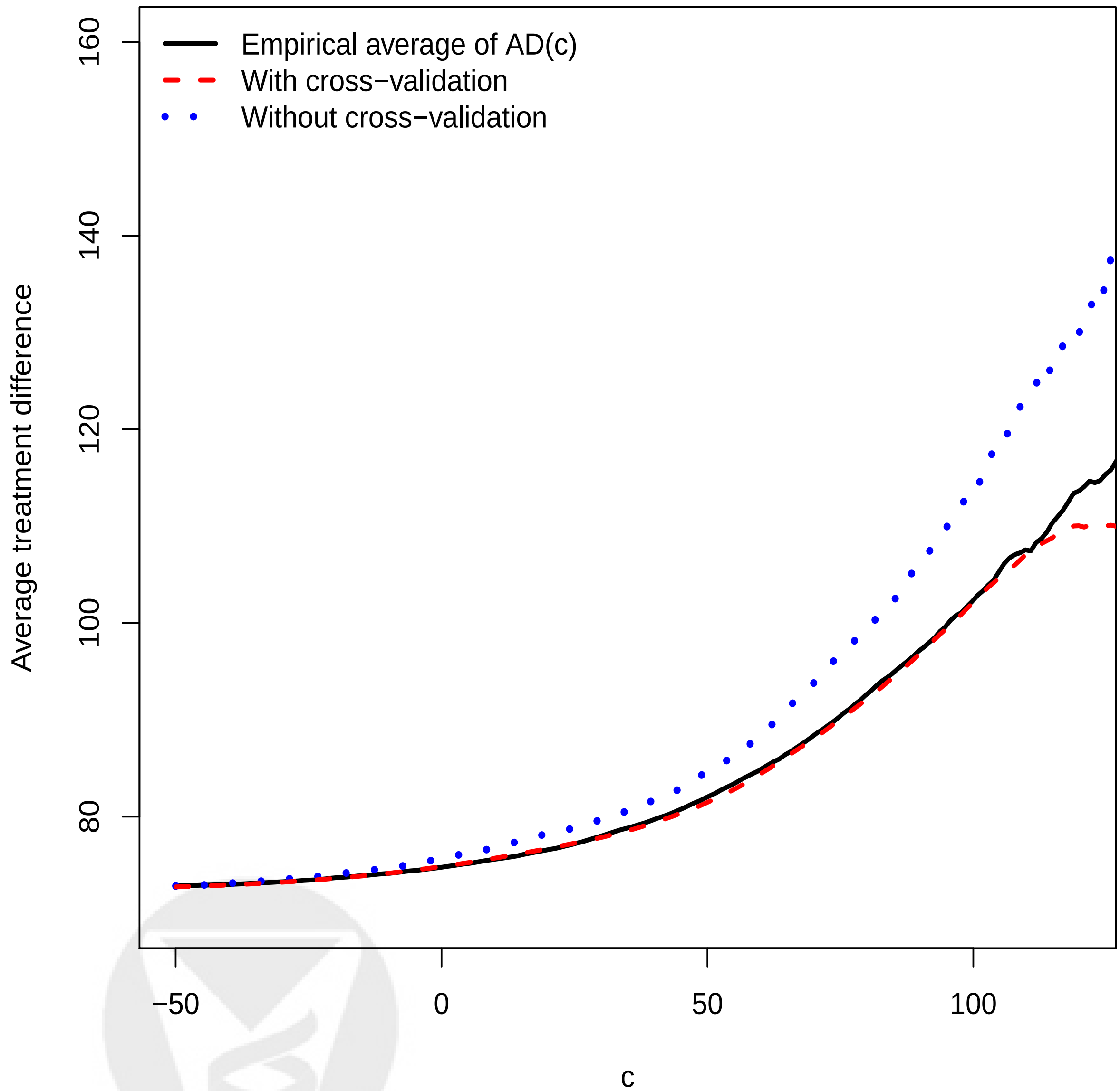

Figure 7: Comparisons between the estimation procedures with and without cross-validation with $n = 838$; the solid curve is the "truth", the dashed curve is the empirical average using cross-validated procedure with a 4:1 ratio of training and evaluation samples, and the dotted curve is the empirical average without using cross-validation.