Harvard University
Harvard University Biostatistics Working Paper Series

Year 2011 *Paper* 132

On the Covariate-adjusted Estimation for an Overall Treatment Difference with Data from a Randomized Comparative Clinical Trial

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

*Stanford University School of Medicine, lutian@stanford.edu
†Harvard University, tcai@hsph.harvard.edu
‡Harvard University, lhzhao@hsph.harvard.edu
**Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper132

On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial

LU TIAN

Department of Health Research & Policy,
Stanford University, Stanford, CA 94305, USA
lutian@stanford.edu

TIANXI CAI

Department of Biostatistics,
Harvard University, Boston, MA 02115, USA

LIHUI ZHAO

Department of Preventive Medicine,
Northwestern University, Chicago, IL 60611, USA

LJ WEI

Department of Biostatistics,
Harvard University, Boston, MA 02115, USA

SUMMARY

To estimate an overall treatment difference with data from a randomized, comparative clinical study, baseline covariates are often utilized to increase the estimation precision. Using the standard analysis of covariance (ANCOVA) technique for making inferences about such an average treatment difference may not be appropriate, especially when the fitted model is non-linear. On the other hand, the novel augmentation procedure recently studied, for example, by Zhang, Davidian and Tsiatis (2008) is quite flexible.

However, in general, it is not clear how to select covariates for augmentation effectively (Shao et al, 2010). An overly adjusted estimator can be severely biased. Furthermore, the results from the standard inference procedure by ignoring the sampling variation from the variable selection process may not be valid. In this paper, we first propose an estimation procedure, which augments the simple treatment contrast estimator directly with covariates. The new proposal is asymptotically equivalent to the aforementioned augmentation method. To select covariates, we utilize the standard lasso procedure. Furthermore, to avoid potential bias of the resulting lasso-type estimator, a cross validation method is used to obtain our final estimation procedure. The validity of the new proposal is justified theoretically and empirically. We illustrate the procedure extensively with a well-known primary biliary cirrhosis clinical trial data set.

Keywords: ANCOVA; Cross validation; Efficiency augmentation; Mayo PBC data; Semi-parametric efficiency

1. INTRODUCTION

For a typical randomized clinical trial to compare two treatments, generally a summary measure θ_0 for quantifying the treatment effectiveness difference can be estimated unbiasedly or consistently using its simple two-sample empirical counterpart, say $\hat{\theta}$. With the subject's baseline covariates, one may obtain a more efficient estimator for θ_0 via a standard analysis of covariance (ANCOVA) technique or a novel augmentation procedure, which is well documented in Zhang, Tsiatis and Davidian (2008) and a series of papers (Leon et al. 2003, Tsiatis 2006, Tsiatis et al. 2008, Lu & Tsiatis 2008, Gilbert et al. 2009, Zhang & Gilbert 2010). The ANCOVA approach can be problematic, especially when the regression model is non-linear, for example, the logistic or Cox model. For this case, the ANCOVA estimator generally does not converge to θ_0, but to a quantity which may be difficult to interpret as a treatment contrast measure. Moreover, in the presence of censored event time observations, this quantity may depend on the censoring

distribution. On the other hand, the above augmentation procedure, referred as ZTD in the literature always produces a consistent estimator for θ_0, provided that the simple estimator $\hat{\theta}$ is consistent.

In theory, the ZTD estimator is asymptotically more efficient than $\hat{\theta}$ no matter how many covariates being augmented. However, in practice, the penalty of an overly augmented estimator can be quite severe. That is, the resulting estimator can be non-trivially biased or its standard error may be larger than that of $\hat{\theta}$. Recently Zhang et al. (2008) showed empirically that the ZTD via the standard stepwise regression for variable selection performs satisfactorily when the number of covariates is not large. In general, however, it is not clear that the standard inference procedures for θ_0 based on estimators augmented by covariates selected via a rather complex variable selection process is appropriate especially when the number of covariates involved is not small relative to the sample size. Therefore, it is highly desirable to develop an estimation procedure to properly and systematically augment $\hat{\theta}$ and make valid inference for the treatment difference based on studies with practical sample sizes.

Now, let Y be the response variable, T be the binary treatment indicator and $\mathbf{Z}$ be a $p-$dimensional vector of covariates or a function thereof including the intercept. The data, $\{(Y_i, T_i, \mathbf{Z}_i), i = 1, \cdots, n\}$, consist of n independent copies of $(Y, T, \mathbf{Z})$, where T and $\mathbf{Z}$ are independent of each other. Let $P(T = 1) = \pi \in (0, 1)$. First, suppose that we are interested in the mean difference: $\theta_0 = E(Y|T = 1) - E(Y|T = 0)$. A simple unadjusted estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{(T_i - \pi)Y_i}{\pi(1-\pi)},$$

which consistently estimates θ_0. To improve efficiency in estimating θ_0, one may employee the standard ANCOVA procedure by fitting the following linear regression *working* model:

$$E(Y|T, \mathbf{Z}) = \theta T + \gamma'\mathbf{Z},$$

where θ and γ are unknown parameters. Since $T \perp \mathbf{Z}$ and $\{(T_i, \mathbf{Z}_i), i = 1, \cdots, n\}$ are independent copies of $(T, \mathbf{Z})$, the resulting ANCOVA estimator is asymptotically equivalent to

$$\hat{\theta} - \hat{\gamma}' \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \right\}, \tag{1.1}$$

where $\hat{\gamma}$ is the ordinary least square estimator for γ of the model $E(Y|\mathbf{Z}) = \gamma'\mathbf{Z}$. The $\hat{\gamma}$ converges to

$$\gamma_0 = \text{argmin}_\gamma E(Y - \gamma'\mathbf{Z})^2,$$

as $n \to \infty$. It follows that the ANCOVA estimator is asymptotically equivalent to

$$\hat{\theta} - \gamma_0' \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \right\}. \tag{1.2}$$

In theory, since $\hat{\theta}$ is consistent to θ_0, the ANCOVA estimator is also consistent to θ_0 and more efficient than $\hat{\theta}$ regardless of whether the above working model is correctly specified. Note that the nonparametric ANCOVA estimator proposed by Koch et al. (1998) and ZTD estimator are also asymptotically equivalent to (1.2), which was noted by Tsiatis et al. (2008). We give details of this equivalence in Appendix A.

The novel ZTD procedure is derived by specifying optimal estimating functions under a very general semi-parametric setting. The efficiency gain from the ZTD has been elegantly justified using the semi-parametric inference theory (Tsiatis 2006). The ZTD is much more flexible than the ANCOVA method in that it can handle cases when the summary measure θ_0 is beyond the simple difference of two group means. On the other hand, the ANCOVA method may only work under the above simple linear regression model.

In this paper, we study the estimator (1.1), which augments $\hat{\theta}$ directly with the covariates. The key question is how to choose $\hat{\gamma}$ in (1.1) especially when p is not small with respect to n. Here, we utilize the lasso procedure with a cross validation process to construct a systematic procedure for selecting co-

variates to increase the estimation precision. The validity of the new proposal is justified theoretically and empirically via an extensive simulation study. The proposal is also illustrated with the data from a clinical trial to evaluate a treatment for a specific liver disease.

2. ESTIMATING THE TREATMENT DIFFERENCE VIA PROPER AUGMENTATION FROM COVARIATES

For a general treatment contrast measure θ_0 and its simple two sample estimator $\hat{\theta}$, assume that

$$\hat{\theta} - \theta_0 = n^{-1} \sum_{i=1}^{n} \tau_i(\eta) + o_p(\frac{1}{\sqrt{n}}),$$

where $\tau_i(\eta)$ is the influence function from the ith observation, η is a vector of unknown parameters, and $i = 1, \cdots, n$. Note that the influence function generally only involves a rather small number of unknown parameters, which is not dependent on $\mathbf{Z}$. Let $\hat{\eta}$ denote the consistent estimator for η. Generally, the above asymptotic expansion is also valid with τ_i being replaced by $\tau_i(\hat{\eta})$. Now, (1.2) can be rewritten as

$$\hat{\theta} - \gamma_0' \left(n^{-1} \sum_{i=1}^{n} \xi_i \right),$$

where $\xi_i = (T_i - \pi)\mathbf{Z}_i / \{\pi(1-\pi)\}, i = 1, \cdots, n$. Then $\hat{\gamma}$ in (1.1) is the minimizer of

$$\sum_{i=1}^{n} (\tau_i(\hat{\eta}) - \gamma' \xi_i)^2. \tag{2.1}$$

When the dimension of $\mathbf{Z}$ is not small, to obtain a stable minimizer, one may consider the following regularized minimand:

$$L_\lambda(\gamma) = \sum_{i=1}^{n} (\tau_i(\hat{\eta}) - \gamma' \xi_i)^2 + \lambda |\gamma|,$$

where λ is the lasso tuning parameter (Tibshirani 1996) and $|\cdot|$ denote the L_1 norm for a vector. For any fixed λ, let the resulting minimizer be denoted by $\hat{\gamma}(\lambda)$. The corresponding augmented estimator and its

variance estimator are

$$\hat{\theta}_{lasso}(\lambda) = \hat{\theta} - \hat{\gamma}(\lambda)' \left(n^{-1} \sum_{i=1}^{n} \xi_i \right)$$

and

$$\hat{V}_{lasso}(\lambda) = n^{-2} \sum_{i=1}^{n} \{\tau_i(\hat{\eta}) - \hat{\gamma}(\lambda)'\xi_i\}^2, \tag{2.2}$$

respectively. When the dimension of $\mathbf{Z}$ is small relative to the sample size, one may ignore the variability of $\hat{\gamma}(\lambda)$ and treat it as a constant when we make inferences about θ_0. In general, however, for practical sample sizes, $\hat{\theta}_{lasso}(\lambda)$ can be substantially biased partly due to the fact that $\hat{\gamma}(\lambda)$ and $\{\xi_i, i = 1, \cdots, n\}$ are correlated. In Appendix B we show via a simple example this undesirable feature of the above estimation procedure.

One possible solution to solve the above problem is to reduce the correlation between $\hat{\gamma}(\lambda)$ and ξ_i using a cross validation procedure. Specifically, we randomly split the data into K non-overlapping sets $\{\mathcal{D}_1, \cdots, \mathcal{D}_K\}$ and construct an estimator for θ_0 :

$$\hat{\theta}_{cv}(\lambda) = \hat{\theta} - \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_{(-i)}(\lambda)'\xi_i,$$

where $i \in \mathcal{D}_{k_i}$, $\hat{\gamma}_{(-i)}(\lambda)$ is the minimizer of

$$\sum_{j \notin \mathcal{D}_{k_i}} (\tau_j(\hat{\eta}_{(-i)}) - \gamma'\xi_j)^2 + \lambda|\gamma|,$$

and $\hat{\eta}_{(-i)}$ is a consistent estimator for η with all data, but not from $\mathcal{D}_{k_i}$. Note that $\hat{\gamma}_{(-i)}(\lambda)$ and ξ_i are independent and no extra bias would be added from $\hat{\theta}_{cv}(\lambda)$ to $\hat{\theta}$. When $n \gg p$, the variance of $\hat{\theta}_{cv}(\lambda)$ can be estimated by $\hat{V}_{lasso}(\lambda)$ given in (2.2). However $\hat{V}_{lasso}(\lambda)$ tends to underestimate its true variance when p is not small.

Here, we utilize the above cross validation procedure to construct a natural variance estimator:

$$\hat{V}_{cv}(\lambda) = n^{-2}\sum_{i=1}^{n}\{\tau_i(\hat{\eta}_{(-i)}) - \hat{\gamma}'_{(-i)}(\lambda)\xi_i\}^2.$$

In Appendix C, we justify that this estimator is better than $\hat{V}_{lasso}(\lambda)$. Moreover, when λ is close to zero and p is large, that is, one almost uses the standard least square procedure to obtain $\hat{\gamma}_{(-i)}(\lambda)$, the above variance estimate can be modified slightly for improving its estimation accuracy (see Appendix C for details). A natural "optimal" estimator using the above lasso procedure is $\hat{\theta}_{opt} = \hat{\theta}_{cv}(\hat{\lambda})$, where $\hat{\lambda}$ is the penalty parameter value, which minimizes $\hat{V}_{cv}(\lambda)$ over a range of λ values of interest.

3. APPLICATIONS

In this section, we show how to apply the new estimation procedure to various cases. To this end, we only need to determine the initial estimate $\hat{\theta}$ for the contrast measure of interest and its corresponding first order expansion in each application. Firstly, we consider the case that the response is continuous or binary and the group mean difference is the parameter of interest. Here

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{(T_i - \pi)Y_i}{\pi(1-\pi)}.$$

In this case, it is straightforward to show that

$$\hat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i(Y_i - \hat{\mu}_1)}{\pi} - \frac{(1-T_i)(Y_i - \hat{\mu}_0)}{1-\pi}\right\} + o_p(\frac{1}{\sqrt{n}}),$$

where $\eta = (\mu_1, \mu_0)'$, $\hat{\mu}_1 = \sum_{i=1}^{n} T_i Y_i/\pi n$, and $\hat{\mu}_0 = \sum_{i=1}^{n}(1-T_i)Y_i/(1-\pi)n$.

Now, when the response is binary with success rate p_j for the treatment group $j, j = 0, 1$, but $\theta_0 = \log\{p_1(1-p_0)/p_0/(1-p_1)\}$, then

$$\hat{\theta} = \log(\hat{p}_1) - \log(1-\hat{p}_1) - \log(\hat{p}_0) + \log(1-\hat{p}_0),$$

where $\hat{p}_1 = \sum_{i=1}^n T_i Y_i/\pi n$, and $\hat{p}_0 = \sum_{i=1}^n (1-T_i)Y_i/(1-\pi)n$. For this case,

$$\hat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(Y_i - \hat{p}_1)T_i}{\pi\hat{p}_1(1-\hat{p}_1)} - \frac{(Y_i-\hat{p}_0)(1-T_i)}{(1-\pi)\hat{p}_0(1-\hat{p}_0)}\right\} + o_p(\frac{1}{\sqrt{n}}).$$

Lastly, we consider the case when Y is the time to a specific event, but may be censored by an independent censoring variable. To be specific, we observe $(\tilde{Y}, \Delta)$ where $\tilde{Y} = Y \wedge C$, $\Delta = I(Y < C)$, C is the censoring time and $I(\cdot)$ is the indicator function. A most commonly used summary measure for quantifying the treatment difference in survival analysis is the ratio of two hazard functions. The two sample Cox estimator is often used to estimate such a ratio. However, when the proportional hazards assumption between two groups is not valid, this estimator converges to a parameter which may be difficult to interpret as a measure of the treatment difference. Moreover, this parameter depends on the censoring distribution. Therefore, it is desirable to use a model-free summary measure for the treatment contrast. One may simply use the survival probability at a given time t_0 as a model-free summary for survivorship. For this case, $\theta_0 = P(Y > t_0|T=1) - P(Y > t_0|T=0)$ and $\hat{\theta} = \hat{S}_1(t_0) - \hat{S}_0(t_0)$, where $\hat{S}_j(\cdot)$ is the Kaplan-Meier estimator of the survival function of Y in group $j, j = 0, 1$. For this case, $\hat{\theta} - \theta_0$

$$= n^{-1}\sum_{i=1}^{n}\left[-\frac{T_i}{\pi}\hat{S}_1(t_0)\int_0^{t_0}\frac{d\hat{M}_{i1}(s)}{\sum_{j=1}^N I(\tilde{Y}_j \geqslant s)T_j} + \frac{1-T_i}{1-\pi}\hat{S}_0(t_0)\int_0^{t_0}\frac{d\hat{M}_{i0}(s)}{\sum_{j=1}^N I(\tilde{Y}_j \geqslant s)(1-T_j)}\right] + o_p(\frac{1}{\sqrt{n}}),$$

where

$$\hat{M}_{ij}(s) = I(\tilde{Y}_i \leqslant s)\Delta_i - \int_0^s I(\tilde{Y}_j \geqslant u)d\left\{T_i\hat{\Lambda}_1(u) + (1-T_i)\hat{\Lambda}_0(u)\right\},$$

and $\hat{\Lambda}_j(\cdot)$ is the Nelson-Alan estimator for the cumulative hazard function of Y in group j (Flemming & Harrington 1991).

To summarize a global survivorship beyond using t-year survival rates, one may use the mean survival time. Unfortunately, in the presence of censoring, such a measure cannot be estimated well. An alternative is to use the so-called restricted mean survival time, that is, the area under the survival function up to time

point t_0. The corresponding consistent estimator is the area under the Kaplan-Meier curve. For this case, $\theta_0 = E(Y \wedge t_0|T=1) - E(Y \wedge t_0|T=0)$ and

$$\hat{\theta} = \int_0^{t_0} \hat{S}_1(s)ds - \int_0^{t_0} \hat{S}_0(s)ds,$$

For this case, $\hat{\theta} - \theta_0$

$$= n^{-1} \sum_{i=1}^{n} \left[-\frac{T_i}{\pi} \int_0^{t_0} \left\{ \frac{\int_s^{t_0} \hat{S}_1(t)dt}{\sum_{j=1}^{N} I(\tilde{Y}_j \geqslant s) T_j} \right\} d\hat{M}_{i1}(s) + \frac{1-T_i}{1-\pi} \int_0^{t_0} \left\{ \frac{\int_s^{t_0} \hat{S}_0(t)dt}{\sum_{j=1}^{N} I(\tilde{Y}_j \geqslant s)(1-T_j)} \right\} d\hat{M}_{i0}(s) \right] + o_p(\frac{1}{\sqrt{n}}).$$

4. A Simulation Study

We conducted an extensive simulation study to examine the finite sample performance of the new estimates $\hat{\theta}_{cv}(\lambda)$ and $\hat{\theta}_{opt}$ for θ_0. Specifically, under various practical settings, we investigate whether $\hat{V}_{cv}(\lambda)$ estimates the true variance of $\hat{\theta}_{cv}(\lambda)$ well. Furthermore, we examine the finite sample properties for the interval estimation procedure based on the optimal $\hat{\theta}_{opt}$. For all cases studied, we find that the proposed estimation procedure performs well. Moreover, although $\hat{V}_{lasso}(\lambda)$ in (2.2) is an asymptotically consistent estimator for the variance of $\hat{\theta}_{cv}(\lambda)$, we find that it can be substantially smaller than the true variance.

As a specific example in our numerical study, we consider the following models for generating the data:

$$Y = \sum_{j=1}^{20} \frac{j}{20} Z_{[j]} + N(0,1), \quad \text{for} \quad T = 0, \text{ and}$$

$$Y = 1 + \sum_{j=1}^{20} \frac{j}{20} Z_{[j]} + N(0,1), \quad \text{for} \quad T = 1,$$

where $(Z_{[1]}, \cdots, Z_{[100]})'$ is generated from the standard multivariate normal distribution. Here, $n = 200, p = 100$ and the group mean difference $\theta_0 = 1$. For each generated data set, the 20-fold cross validation is used to calculate $\hat{\theta}_{cv}(\lambda)$ and $\hat{V}_{cv}(\lambda)$ over a sequence of tuning parameters $\{\lambda_1, \lambda_2, \cdots, \lambda_{100}\}$,

where λ_1 is chosen such that $\hat{\gamma}(\lambda_1) = 0$ for all the simulated data sets, $\{\lambda_1, \cdots, \lambda_{99}\}$ is a sequence of values evenly decreasing from λ_1 to $\lambda_{99} = 10^{-3}\lambda_1$ on the log scale, and $\lambda_{100} = 0$. In figure 1(a), we present the empirical average for $\hat{V}_{cv}(\lambda)$ (blue curve) and the empirical variance of $\hat{\theta}_{cv}(\lambda)$ (red curve) based on 5000 replications, where the x-axis is the order of those 100 λ values. The empirical average of $\hat{V}_{lasso}(\lambda)$ is also presented (green curve). The figure shows on average $\hat{V}_{cv}(\lambda)$ is almost identical to the empirical variance of $\hat{\theta}_{cv}(\lambda)$. On the other hand, $\hat{V}_{lasso}(\lambda)$ without using cross validation tends to substantially under estimate the true variance.

In this same set of simulation, we also generate a binary response Y from the following logistic regression model

$$\mathrm{P}(Y = 1|T = 1) = \frac{\exp\{1 + \sum_{j=1}^{20} \frac{j}{20} Z_{[j]}\}}{1 + \exp\{1 + \sum_{j=1}^{20} \frac{j}{20} Z_{[j]}\}}, \text{ and}$$

$$\mathrm{P}(Y = 1|T = 0) = \frac{\exp\{\sum_{j=1}^{20} \frac{j}{20} Z_{[j]}\}}{1 + \exp\{\sum_{j=1}^{20} \frac{j}{20} Z_{[j]}\}}.$$

Here, $n = 200, p = 100$ and the $\log(\text{odds ratio})$ is the parameter of interest. The results on variance estimates are shown in Figure 1(b). Again, the variance estimator $\hat{V}_{cv}(\lambda)$ behaves well, but not $\hat{V}_{lasso}(\lambda)$.

Lastly, we simulate the survival time from the following Cox regression model

$$Y = \epsilon_0 \exp\left\{1 + \sum_{j=1}^{20} \frac{j}{20} Z_{[j]}\right\}, \quad \text{for} \quad T = 1, \text{ and}$$

$$Y = \epsilon_0 \exp\left\{\sum_{j=1}^{20} \frac{j}{20} Z_{[j]}\right\}, \quad \text{for} \quad T = 0,$$

where ϵ_0 follows the unit exponential distribution. The censoring distribution is generated from $U(0, 3)$, which yields approximately 50% of censoring. Here, $n = 200, p = 100$ and the difference in mean survival time truncated at $t_0 = 2.2$ is the parameter of interest. The simulation results on variance estimates are shown in Figure 1(c). The $\hat{V}_{cv}(\lambda)$ curve has almost no any meaningful difference from the "true"

variance curve of $\hat{\theta}_{cv}(\lambda)$. Note that for the above three sets of simulation, we also find that $\hat{\theta}_{cv}(\lambda)$ almost has no bias for estimating θ_0 as expected.

We also examine the performance of the optimal estimator $\hat{\theta}_{opt} = \hat{\theta}_{cv}(\hat{\lambda})$, where $\hat{\lambda}$ is chosen to be the minimizer of $\hat{V}_{cv}(\lambda), \lambda \in \{\lambda_1, \cdots, \lambda_{100}\}$. First, with the above 5000 simulated data, one can obtain the empirical variances of $\hat{\theta}_{cv}(\lambda)$. Let λ_0 be the minimizer of the curve of such empirical variances over λ. Then, for each set of the above simulation, we generated 5000 0.95 confidence intervals based on $\hat{\theta}_{opt}$ and $\hat{V}_{opt} = \hat{V}_{cv}(\hat{\lambda})$. We compute the corresponding empirical coverage level and the length. For comparisons, we also obtain those values based on the simple estimator $\hat{\theta}$ and its variance estimate $\hat{V}$, and also based on the cross validation estimation procedure with λ_0. The results are summarized in Table 1. The coverage levels for $\hat{\theta}_{opt}$ are close to the nominal counterparts and the interval lengths are almost identical to those for the estimate with the true optimal λ_0. On the other hand, the simple estimate $\hat{\theta}$ tends to have substantial wider interval estimates that $\hat{\theta}_{opt}$.

For all cases studied, the estimate $\hat{\theta}_{opt}$ is almost unbiased and can substantially improve the efficiency of the simple estimate $\hat{\theta}$ for the overall treatment difference in terms of narrowing the length of the confidence interval of θ_0. Furthermore, the variability in $\hat{\lambda}$ is almost negligible in making inference for θ_0 based on $\hat{\theta}_{cv}(\hat{\lambda})$.

5. An Example

We illustrate the new proposal with the data from a clinical trial to compare D-penicillmain and placebo for patients with primary biliary cirrhosis (PBC) of liver (Therneau & Grambsch 2000). The primary endpoint is the time to death. The trial was conducted between 1974 and 1984. For illustration, we use the difference of two restricted mean survival times up to $t_0 = 3650$ (days) as the primary parameter θ_0 of interest.

Moreover, we consider 18 baseline covariates for augmentation: gender, stage (1, 2, 3, and 4), presence of ascites, edema, hepatomegaly or enlarged liver, blood vessel malformations in the skin, log-transformed age, serum albumin, alkaline phosphotase, aspartate aminotransferase, serum bilirubin, serum cholesterol, urine copper, platelet count, standardized blood clotting time and triglycerides. There are 276 patients with complete covariate information (136 and 140 in control and D-penicillmain arms, respectively). The data used in our analysis are given in the Appendix D.1 of Flemming & Harrington (1991). Figure 2(a) provides the Kaplan-Meier curves for the two treatment groups. The simple two sample estimate $\hat{\theta}$ is 115.2 (days) with an estimated standard error $\hat{V}$ of 156.6 (days). The corresponding 95% confidence interval for the difference is (-191.8, 422.1) (days). The optimal estimate $\hat{\theta}_{opt}$ augmented additively with the above 18 coavariates is 106.3 with an estimated standard error $\hat{V}_{opt}$ of 121.4. These estimates were obtained via a 23-fold cross validation (Note that $276 = 23 \times 12$) described in Section 2. The corresponding 95% confidence interval is (-131.8, 344.4).

To examine how robust the new proposal is with respect to different augmentations. We consider a case which includes the above 18 covariates, but also their quadratic terms as well as all their two-way interactions. The dimension of $\mathbf{Z}$ is 178 for this case. The resulting optimal $\hat{\theta}_{opt}$ is 110.1 with an estimated standard error of 122.6. Note the resulting estimates are amazingly close to those based on the augmented procedure with 18 covariates only.

To examine the advantage of using the cross validation for the standard error estimation, in Figure 2(b), we plot $\hat{V}_{cv}(\lambda)$ and $\hat{V}_{lasso}(\lambda)$ over the order of 100 λ's, which were generated using the same approach as in Section 4. Note that $\hat{V}_{lasso}(\lambda)$ is substantially smaller than $\hat{V}_{cv}(\lambda)$, especially when λ approaches to 0, that is, there is no penalty for the L_2 loss function. For $\hat{\theta}_{opt}$, $\hat{V}_{lasso}$ is about 20% smaller than its cross validated counterpart.

It has been shown via numerical studies that the ZTD performs well via the standard stepwise regression by ignoring the sampling variation of the estimated weights when the dimension of $\mathbf{Z}$ is not large with respect to n. However, it is not clear how the ZTD augmentation performs with a relatively high dimensional covariate vector $\mathbf{Z}$. It would be interesting to compare the ZTD and the new proposal with the PBC data. To this end, we implement ZTD augmentation procedure using (1) baseline covariates $(p = 18)$; (2) baseline covariates and their quadratic transformations as well as all their two-way interactions $(p = 178)$; and (3) only five baseline covariates: edema and log-transformed age, serum albumin, serum bilirubin and standardized blood clotting time, which were selected in building a multivariate Cox regression model to predict the patient's survival by Therneau & Grambsch (2000). Note that the ZTD procedure augments the following estimating equations for θ_0

$$\sum_{i=1}^{n} \frac{(1-T_i)\tilde{\Delta}_i}{\hat{K}_0(\tilde{Y}_i \wedge t_0)}[\tilde{Y}_i \wedge t_0 - a_{t_0}] = 0,$$

$$\sum_{i=1}^{n} \frac{T_i\tilde{\Delta}_i}{\hat{K}_1(\tilde{Y}_i \wedge t_0)}[\tilde{Y}_i \wedge t_0 - a_{t_0} - \theta] = 0,$$

where a_{t_0} is the restricted mean for the comparator and θ is the treatment difference, $\tilde{\Delta}_i = I(Y_i \wedge t_0 < C_i)$ and $\hat{K}_j(\cdot)$ is the Kaplan-Meier estimate for the survival function of censoring time C in group $T = j, j = 0, 1$. In Table 2, we present the resulting ZDT point estimates and their corresponding standard error estimates for the above three cases. We used the standard forward stepwise regression procedure to select the augmentation covariates with the entry Type I error rate of 0.10 (Zhang et al., 2008; Zhang & Gilbert, 2010). It appears that using the entire data set for selecting covariates and making inferences about θ_0 may introduce nontrivial bias and an overly optimistic standard error estimate when p is large. On the other hand, the new procedure does not lose efficiency and yields similar result as ZTD procedure when p is small.

6. REMARKS

The new proposal performs well even when the dimension of the covariates involved for augmentation is not large. The new estimation procedure may be implemented for improving estimation precision regardless of the marginal distributions of the covariate vectors between two treatment groups being balanced. On the other hand, to avoid post ad hoc analysis, we strongly recommend that the investigators prespecify the set of all potential covariates for adjustment in the protocol or the statistical analysis plan before the data from the clinical study are unblinded.

The stratified estimation procedure for the treatment difference is also commonly used for improving the estimation precision using baseline covariate information. Specifically, we divide the population into K strata based on baseline variables, denoted by $\{\mathbf{Z} \in B_1\}, \cdots, \{\mathbf{Z} \in B_K\}$, the stratified estimator is

$$\hat{\theta}_{str} = \frac{\sum_{k=1}^{K} \hat{\theta}_k w_k}{\sum_{k=1}^{K} w_k},$$

where $\hat{\theta}_k$ and w_k are corresponding simple two sample estimator for the treatment difference and the weight for the kth stratum, $k = 1, \cdots, K$. In general, the underlying treatment effect may vary across strata and consequently the stratified estimator may not converge to θ_0. If θ_0 is the mean difference between two groups and w_k is the size of the kth stratum, $\hat{\theta}_{str}$ is a consistent estimator for θ_0. Like the ANCOVA, the stratified estimation procedure may be problematic. On the other hand, one may use the indicators $\{I(\mathbf{Z} \in B_1), \cdots, I(\mathbf{Z} \in B_K)\}'$ to augment $\hat{\theta}$ to increase the precision for estimating the treatment difference θ_0.

In this paper we follow the novel approach taken, for example, by Zhang et al. (2008) for augmenting the simple two sample estimator, but present a systematic, practical procedure for choosing covariates for making valid inferences about the overall treatment difference. When p is large, there are several ad-

vantages over other approaches for augmenting $\hat{\theta}$ with covariates. Firstly, it avoids the complex variable selection step in two arms separately as proposed in Zhang et al. (2008). Secondly, compared with other variable selection methods such as the stepwise regression, the lasso method directly controls the variability of $\hat{\gamma}$, which is important to ensure the validity of the statistical inference for the treatment difference. When λ increases from 0 to $+\infty$, the resulting estimator varies from the fully augmented estimator using all the components of $\mathbf{Z}_i$ to $\hat{\theta}$. The lasso procedure also possesses superior computational efficiency with high dimensional covariates to alternatives. Lastly, since the ZTD estimator can also be viewed as a generalized method of moment estimator with

$$\begin{pmatrix} \theta - \hat{\theta}_0 \\ n^{-1}\sum_{i=1}^{n} \xi_i \end{pmatrix} \approx 0$$

as moment conditions (Hall 2005), the cross validation method introduced here may be extended to a much broader context than the current setting.

It is important to note that if a permuted block treatment allocation rule is used for assigning patients to the two treatment groups, the augmentation method proposed in the paper can be easily modified. For instance, for the K-fold cross validation process, one may choose the sets $\{\mathcal{D}_k, k = 1, \cdots, K\}$ so that each permuted block would not be in different sets.

For assigning patients to the treatment groups, a stratified random treatment allocation rule is also often utilized to ensure a certain level of balance between the two groups in each stratum. For this case, a weighted average θ_0 of the treatment differences θ_{k0} with weight $w_k, k = 1, \cdots, K$, across K strata may be the parameter of interest for quantifying an overall treatment contrast. Let $\hat{\theta}_k$ be the simple two sample estimator for θ_{k0} and $\hat{w}_k$ be the corresponding empirical weight for w_k. Then the weight average $\hat{\theta} = \sum_k \hat{w}_k \hat{\theta}_k / \sum_k \hat{w}_k$ is the simple estimator for θ_0. For the kth stratum, one may use the same approach as discussed in this paper to augment $\hat{\theta}_k$, let the resulting optimal estimator be denoted by $\hat{\theta}_{opt,k}$. Then we

can use the weighted average $\sum_k \hat{w}_k \hat{\theta}_{opt,k} / \sum_k \hat{w}_k$ to estimate θ_0. On the other hand, for the case with the dynamic treatment allocation rules (see, for example, (Pocock & Simon 1975)), it is not clear how to obtain a valid variance estimate even for the simple two sample estimator $\hat{\theta}$ (Shao et al. 2010). How to extend the augmentation procedure to cases with more complicated treatment allocation rule warrants further research.

7. APPENDIX

7.1 *Appendix A: Asymptotical Equivalence Between ZTD and ANCOVA.*

When the group mean is the parameter of interest, the naive estimator for θ_0 can viewed as the root of the estimating equation

$$\sum_{i=1}^{n} \begin{pmatrix} T_i \\ 1-T_i \end{pmatrix} S_0(\theta, a, Y_i, T_i) = \sum_{i=1}^{n} \begin{pmatrix} T_i \\ 1-T_i \end{pmatrix} (Y_i - a - T_i\theta) = 0,$$

where $a = E(Y|T=0)$ is a nuisance parameter. In the ZTD augmentation procedure, one may augment this simple estimating equation via following steps

- Obtain the initial estimator

$$\begin{pmatrix} \hat{\theta} \\ \hat{a} \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \frac{(T_i-\pi)Y_i}{\pi(1-\pi)} \\ \frac{(1-T_i)Y_i}{1-\pi} \end{pmatrix}$$

  from the original estimating equation

- Obtain $\hat{\beta}_1$ and $\hat{\beta}_0'$ by minimizing the objective function

$$\sum_{i=1}^{n} T_i \{S_0(\hat{\theta}, \hat{a}, Y_i, T_i) - \beta_1' \mathbf{Z}_i\}^2$$

  and

$$\sum_{i=1}^{n} (1-T_i) \{S_0(\hat{\theta}, \hat{a}, Y_i, T_i) - \beta_0' \mathbf{Z}_i\}^2$$

respectively. In other words, using $\hat{\beta}_j'\mathbf{Z}$ to approximate $E\{S_0(\theta_0, a_0; Y, T)|\mathbf{Z}, T = j\}$.

- Solve the augmented estimating equations

$$\sum_{i=1}^{n} \begin{pmatrix} T_i \\ 1 - T_i \end{pmatrix} S_0(\theta, a, Y_i, T_i) - \sum_{i=1}^{n} (T_i - \pi) \begin{pmatrix} \hat{\beta}_1'\mathbf{Z}_i \\ \hat{\beta}_0'\mathbf{Z}_i \end{pmatrix} = 0$$

  to obtain the ZTD estimator.

The resulting ZTD estimator is always asymptotically more efficient than the naive counterpart and a simple sandwich variance estimator can be used to consistently estimate the variance of the new estimator. It has been shown that ZTD estimator is asymptotically the most efficient one from the class of the estimators

$$\mathcal{A} = \left\{ \hat{\theta}_\gamma = \hat{\theta} - \gamma' \left\{ n^{-1} \sum_{i=1}^{n} \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \right\} \,\middle|\, \gamma \in R^p \right\},$$

whose members are all consistent for θ_0 and asymptotically normal. Since

$$\hat{\theta} - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(T_i - \pi)Y_i}{\pi(1-\pi)} - \theta_0 \right\},$$

the optimal weight minimizing the variance of

$$\hat{\theta} - \gamma' \frac{1}{n} \sum_{i=1}^{n} \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)}$$

is simply

$$\left[E\left\{ \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \right\}^{\otimes 2} \right]^{-1} E\left[\frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \left\{ \frac{(T_i - \pi)Y_i}{\pi(1-\pi)} - \theta_0 \right\} \right]$$
$$= [E(\mathbf{Z}_i^{\otimes 2})]^{-1} E(\mathbf{Z}_i Y_i) = \gamma_0$$

Therefore, ZTD estimator is asymptotically equivalent to the commonly used ANCOVA estimator. This equivalence is noted in Tsiatis et al. (2008).

7.2 *Appendix B: An Example on the Potential Bias of ZTD Procedure*

In this section we show via an example that if we estimate the weights γ with the entire data set and then construct $\hat{\theta}_{lasso}(0)$ and ZTD estimator with the same data set, the resulting estimation procedure can be substantially biased. To this end, we consider the following models to generate the data:

$$Y = \sum_{j=1}^{10} \frac{1 - Z_{[j]}^2}{2} - \sum_{j=1}^{20} \frac{j}{20} Z_{[j]} + N(0,1), \quad \text{for} \quad T = 0, \text{ and}$$
$$Y = \sum_{j=1}^{10} \frac{Z_{[j]}^2 - 1}{2} + \sum_{j=1}^{20} \frac{j}{20} Z_{[j]} + N(0,1), \quad \text{for} \quad T = 1,$$

where $(Z_{[1]}, \cdots, Z_{[20]})'$ is the 20-dimensional standard multivariate normal. We let the total sample size be 200 with 1:1 random allocations. Here, the true parameter $\theta_0 = 0$. We construct $\hat{\theta}_{lasso}(0)$ and the ZTD estimator. Based on 5000 simulated data sets from the above model, we obtain the average bias and the standard error. The empirical bias of $\hat{\theta}_{lasso}(0)$ is -0.09, which is 24% of the empirical standard error. The ZTD estimator is slightly more biased than $\hat{\theta}_{lasso}(0)$ with an empirical bias of -0.12, 31% of its empirical standard error.

7.3 *Appendix C: Justification of the cross validation based variance estimator for* $\hat{\theta}_{cv}(\lambda)$

To justify the cross validation based variance estimator, first consider the expansion

$$\hat{\theta}_{cv}(\lambda) = \left\{ \hat{\theta} - \gamma_0' \left(n^{-1} \sum_{i=1}^{n} \xi_i \right) \right\} - n^{-1} \sum_{i=1}^{n} \{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}' \xi_i.$$

The variance of $\hat{\theta}_{cv}(\lambda)$ can be expressed as $V_{11} + V_{22} - 2V_{12}$, where

$$V_{11} = E \left\{ \hat{\theta} - \gamma_0' \left(n^{-1} \sum_{i=1}^{n} \xi_i \right) \right\}^2,$$

$$V_{22} = \frac{1}{n^2} E \left[\sum_{i=1}^{n} \{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}' \xi_i \right]^2,$$

and

$$V_{12} = \frac{1}{n} E\left[\left\{\hat{\theta} - \gamma_0'\left(n^{-1}\sum_{i=1}^{n}\xi_i\right)\right\}\sum_{i=1}^{n}\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i\right].$$

Firstly

$$\begin{aligned}
V_{12} =& \frac{1}{n^2} E\left[\sum_{i=1}^{n}(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\sum_{i=1}^{n}\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i\right] \\
\approx& \frac{1}{n^2}\sum_{i\neq j} E\left[(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\{\hat{\gamma}_{(-j)}(\lambda) - \gamma_0\}'\right] E\xi_j + \frac{1}{n^2}\sum_{i=1}^{n} E\left[(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i\right] \\
\approx& \frac{1}{n^2}\sum_{i=1}^{n} E\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}' E\left[(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\xi_i\right] \approx 0.
\end{aligned}$$

Therefore, the variance of the augmented estimator $\hat{\theta}_{cv}(\lambda)$ is approximately

$$\begin{aligned}
& V_{11} + V_{22} \\
=& \frac{1}{n}\left[E\{(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)^2\} + E\{(\hat{\gamma}_{(-i)}(\lambda) - \gamma_0)'\xi_i\}^2\right] + \frac{(n-1)}{n} E[\xi_1'\{\hat{\gamma}_{(-1)}(\lambda) - \gamma_0\}\xi_2'\{\hat{\gamma}_{(-2)}(\lambda) - \gamma_0\}] \\
\approx& \hat{V}_{cv}(\lambda) + \frac{(n-1)}{n} E[\xi_1'\hat{\gamma}_{(-1)}(\lambda)\xi_2'\hat{\gamma}_{(-2)}(\lambda)].
\end{aligned}$$

In our experience, $d(\lambda) = E[\xi_1'\hat{\gamma}_{(-1)}(\lambda)\xi_2'\hat{\gamma}_{(-2)}(\lambda)] = O(n^{-2})$ is very small compared with $\hat{V}_{cv}(\lambda) = O(n^{-1})$ and is negligible, when λ is not close zero . Therefore, in general, $\hat{V}_{cv}(\lambda)$ serves as a satisfactory estimator for the variance of $\hat{\theta}_{cv}(\lambda)$. For small λ, to explicitly estimate $d(\lambda)$, the covariance between $\xi_1'\hat{\gamma}_{(-1)}(\lambda)$ and $\xi_2'\hat{\gamma}_{(-2)}(\lambda)$, one may use

$$\hat{d}(\lambda) = \frac{2(K^2-1)}{n(n-1)K}\sum_{1\leqslant i<j\leqslant n}\xi_i'\left\{\frac{K-1}{K}\hat{\gamma}_{(-j)}(\lambda) - \hat{\gamma}(\lambda)\right\}\xi_j'\left\{\frac{K-1}{K}\hat{\gamma}_{(-i)}(\lambda) - \hat{\gamma}(\lambda)\right\} \tag{7.1}$$

as an ad-hoc jackknife-type estimator, where $\hat{\gamma}(\lambda)$ is the lasso solution based on the entire data set. To justify the approximation, first note that when λ is close to zero,

$$\hat{\gamma}(\lambda) - \gamma_0 \approx \sum_{i=1}^{n}\Upsilon_i \quad \text{and} \quad \hat{\gamma}_{(-i)}(\lambda) - \gamma_0 \approx \frac{K}{K-1}\sum_{i\notin\mathcal{D}_{k_i}}\Upsilon_i$$

where Υ_i is the mean zero influence function from the ith observation for $\hat{\gamma}(\lambda)$. Therefore,

$$d(\lambda) = E[\xi_1' \hat{\gamma}_{(-1)}(\lambda) \xi_2' \hat{\gamma}_{(-2)}(\lambda)] \approx \left(1 - \frac{1}{K^2}\right) E[\xi_1' \Upsilon_2 \xi_2' \Upsilon_1],$$

which can be approximated by $\hat{d}(\lambda)$ and one may use $\hat{V}_{cv}(\lambda) + (n-1)\hat{d}(\lambda)/n$ as the variance estimator for the augmented estimator. Note that the difference between $\hat{V}_{cv}$ and its modified version appears to be negligible in all the numerical studies presented in the paper.

REFERENCES

Flemming, T. & Harrington, D. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

Gilbert, P. B., Sato, M. & Sun, X.and Mehrotra, D. V. (2009), 'Efficient and robust method for comparing the immunogenicity of candidate vaccines in randomized clinical trials.', *Vaccine* **27**, 396–401.

Hall, A. (2005), *Generalized Method of Moments (Advanced Texts in Econometrics).*, Oxford University Press, London.

Koch, G., Tangen, C., Jung, J. & Amara, I. (1998), 'Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them.', *Statistics in Medicine* **17**, 1863–1892.

Leon, S., Tsiatis, A. & Davidian, M. (2003), 'Semiparametric efficiency estimation of treatment effect in a pretest-posttest study.', *Biometrics* **59**, 1046–1055.

Lu, X. & Tsiatis, A. (2008), 'Improving efficiency of the log-rank test using auxiliary covariates.', *Biometrika* **95**, 676–694.

Pocock, S. & Simon, R. (1975), 'Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial.', *Biometrics* **31**, 102–115.

Shao, J., Yu, X. & Zhong, B. (2010), 'A theory for testing hypotheses under covariate-adaptive randomization.', *Biometrika* **97**, 347–360.

Therneau, T. & Grambsch, P. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

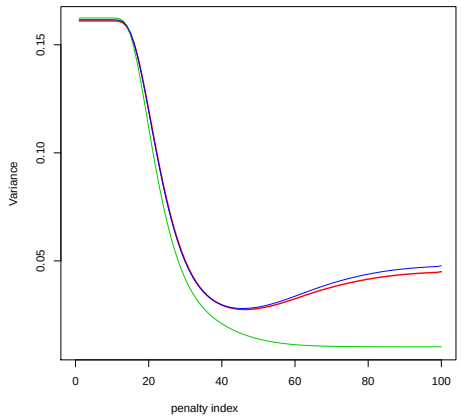
Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tsiatis, A. (2006), *Semiparametric Theory and Missing Data.*, Springer, New York.

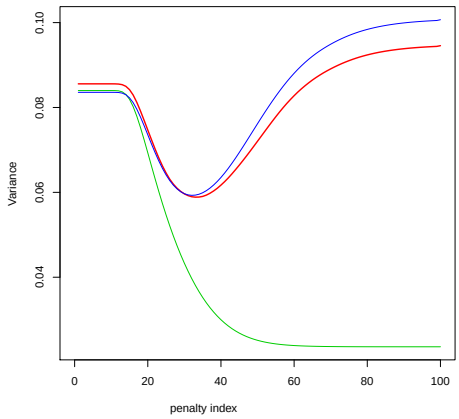
Tsiatis, A., Davidian, M., Zhang, M. & Lu, X. (2008), 'Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach.', *Statistics in Medicine* **27**, 4658–4677.

Zhang, M. & Gilbert, P. B. (2010), 'Increasing the efficiency of prevention trials by incorporating baseline covariates.', *Stat Commun Infect Dis.* **2**, doi:10.2202/1948–4690.1002.

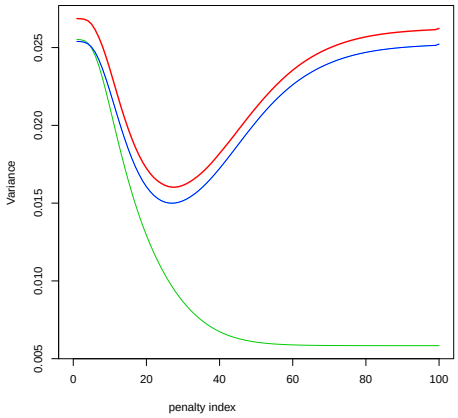
Zhang, M., Tsiatis, A. & Davidian, M. (2008), 'Improving efficiency of inferences in randomized clinical trials using auxiliary covariates.', *Biometrics* **64**, 707–715.



(a) Gaussian outcome

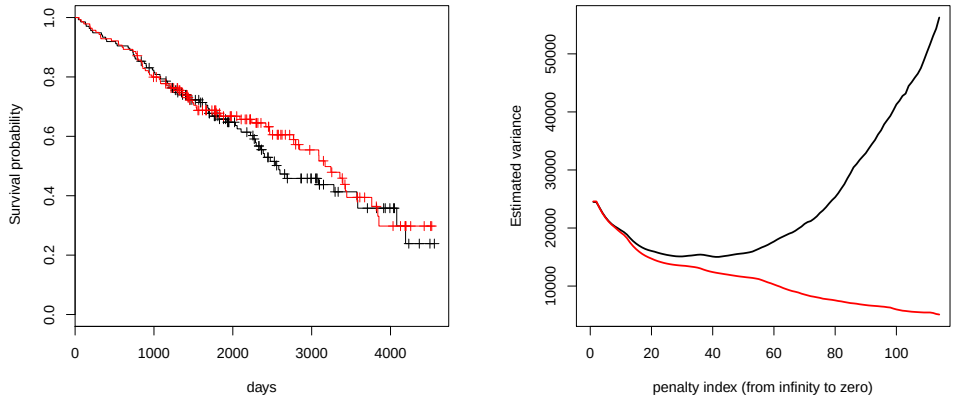


(b) Binary outcome



(c) Survival outcome

Fig. 1. Comparing various estimates for $\hat{\theta}_{cv}(\lambda)$ at $\{\lambda_1, \cdots, \lambda_{100}\}$: the empirical variance of $\hat{\theta}_{cv}(\lambda)$ (red curve); $\hat{V}_{cv}(\lambda)$ (blue curve); $\hat{V}_{lasso}(\lambda)$ (green curve)



(a) Estimated survival functions of D-penicillmain (red) and placebo arms (black)

(b) $\hat{V}_{cv}(\lambda)$ (black) vs. $\hat{V}_{lasso}(\lambda)$ (red)

Fig. 2. Analysis results for primary biliary cirrhosis data

Table 1. The empirical coverage levels and lengths for the 0.95 interval estimation procedure based on $\hat{\theta}_{opt}$ and $\hat{V}_{opt}$ (EAL: empirical length; ECL: empirical coverage level)

	$\hat{\theta}_{opt}$		$\hat{\theta}$		$\hat{\theta}_{cv}(\lambda_0)$	
Response	EAL	ECL	EAL	ECL	EAL	ECL
Continuous	0.644	94.4%	1.578	95.4%	0.653	94.4%
Binary	0.946	94.7%	1.136	94.7%	0.954	94.9%
Survival	0.476	93.8%	0.626	94.4%	0.479	94.1%

Table 2. Comparisons between the new and ZTD estimate with the data from the Mayo Clinic Primary Biliary Cirrhosis clinical trial (SE: estimated standard error)

	The new optimal procedure		ZTD	
p	Estimate	SE	Estimate	SE
5	92.0	121.5	96.3	119.4
18	106.3	121.4	126.4	111.7
178	110.1	122.6	65.3	114.6