Harvard University

Harvard University Biostatistics Working Paper Series

Year 2010 *Paper* 115

Utilizing the Integrated Difference of Two Survival Functions to Quantify the Treatment Contrast for Designing, Monitoring and Analyzing a Comparative Clinical Study

Lihui Zhao[*] Lu Tian[†] Hajime Uno[‡]

Scott D. Solomon[**] Marc A. Pfeffer[††]

J. S. Schindler[‡‡] L. J. Wei[§]

[*] Harvard University, lhzhao@hsph.harvard.edu

[†] Stanford University School of Medicine, lutian@stanford.edu

[‡] Harvard University, huno@hsph.harvard.edu

[**] Brigham and Women's Hospital

[††] Brigham and Women's Hospital

[‡‡] Merck Research Laboratories

[§] Harvard University, wei@hsph.harvard.edu

UTILIZING THE INTEGRATED DIFFERENCE OF TWO SURVIVAL FUNCTIONS TO QUANTIFY THE TREATMENT CONTRAST FOR DESIGNING, MONITORING AND ANALYZING A COMPARATIVE CLINICAL STUDY

L. Zhao[1], L. Tian[2], H. Uno[1,3], S. D. Solomon[4], M. A. Pfeffer[4], J. S. Schindler[5], and L. J. Wei[1,*]

[1]Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A

[2]Department of Health Research and Policy, Stanford University School of Medicine,
Stanford, CA 94305, U.S.A.

[3]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute,
Boston, MA 02115, U.S.A.

[4]Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115, U.S.A.

[5]Merck Research Laboratories, 126 E. Lincoln Avenue, Rahway, NJ 07065, U.S.A.

**email:* wei@hsph.harvard.edu

Summary: In a comparative clinical study with an event time as the endpoint, we often design and monitor the trial using an event-driven approach. That is, the sample size estimate and the timing of the interim reviews are based on the expected and observed numbers of events over time. Estimates of the proportional hazards or differences of two survival functions at specific time points may then be used to quantify the overall treatment contrast. For superiority trials, each of these two estimation procedures has its own merit. However, in this paper, we show that under an equivalence or non-inferiority study setting, when the event rates are low, using estimates of an *average* difference of survival rates over a time interval to design, monitor and analyze the study is a much better strategy than using event-driven based tests or estimates. The choice of this time interval depends on the questions to be answered from the study. We illustrate our proposal using the data from a cardiovascular clinical trial. A numerical study is also conducted to examine the performance of the new proposal.

Key words: Equivalence study; Event driven study; Kaplan-Meier curve; Non-inferiority trial; Post-market study; Proportional hazards estimate.

1. INTRODUCTION

To assess the relative efficacy or safety of two treatments with an event time as the outcome variable in a comparative clinical trial, one generally uses an *expected* or *estimated* number of events over study time to determine the sample size and monitoring schedule. For interim and final data analyses of such an event-driven trial, we typically summarize the results with a plot of two Kaplan-Meier (KM) curves, $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$, the p-value of the logrank test and the proportional hazards (PH) point and interval estimates of an "average" of hazard ratios over the study duration (Cox, 1972). The KM estimates are simply descriptive statistics. The Cox estimate provides a single summary for quantifying the treatment difference. When the proportional hazards (PH) assumption is non-trivially violated, the interpretation of the treatment contrast from the KM curves can be quite different from that of the aforementioned hazard ratio estimates. On the other hand, one may compare two survival functions directly with the differences $\hat{D}(t) = \hat{S}_2(t) - \hat{S}_1(t)$ evaluated at a set of specific time points t's or for all $t \in [t_0, t_1]$, a fixed time interval. For instance, comparisons may be made via pointwise or simultaneous confidence interval estimates for the difference of two survival functions over $[t_0, t_1]$ (Parzen, Wei and Ying, 1997). Moreover, one may consider an integrated (average) difference $\hat{D}$ of $\{\hat{D}(t), t \in [t_0, t_1]\}$ as a summary statistic to quantify the treatment contrast, where

$$\hat{D} = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \hat{D}(t) dt.$$

Such an integrated difference or a weighted version thereof has been proposed as a test statistic for testing the equality of two survival curves, for example, by Pepe and Fleming (1989, 1991). They showed that, under a superiority trial setting, this type of test performs well even under the proportional hazards alternative and can be better than the logrank test when the PH assumption is not valid.

From the estimation point of view, if the difference of two survival functions is approximately equal to a constant over $[t_0, t_1]$, $\hat{D}$ or its weighted version would consistently estimate such a constant. In practice, the assumption of the proportional hazards or a constant survival rate difference over $[t_0, t_1]$ is rarely valid. The Cox estimate or $\hat{D}$ has its own merit for measuring an *overall* treatment difference. However, when we are interested in the treatment difference with respect to relatively long term survival, $\hat{D}$ may provide more relevant information than the PH estimate by choosing an appropriate time interval $[t_0, t_1]$. Moreover, when the event rates are low for both groups, the variance estimate of the PH estimator almost entirely depends on the total number of observed events. This generally leads to a conclusion that the trial is futile due to lack of information. That is, at the end of the study, we cannot say whether the two groups are equivalent or one treatment is superior or non-inferior to the other. To illustrate this point, let us use the data from a clinical trial "Prevention of Events with Angiotensin Converting Enzyme Inhibition (PEACE)." This trial was designed to study whether the ACE inhibitors (ACEi) would be effective for reducing certain cardiovascular related events (Braunwald et al., 2004). In this study, 4158 and 4132 patients were randomly assigned to the trandolapril treatment and placebo arms, respectively, and they underwent randomization from November 1996 to June 2000. The median follow-up time is 4.8 years and the longest is 7 years. For illustration, let the primary event of interest be the death from all causes and consider a subgroup of patients who are younger than 65. Figure 1 gives us the Kaplan-Meier curves based on survival data collected at the end of the study from this subgroup of patients (2119 treated and 2085 controls). There are 99 and 94 events in the control and treated arms, respectively. Except for the unstable tail parts, visually there is no difference between the two curves. The PH point estimate and the corresponding 0.95 confidence interval estimate are 0.97 and (0.73, 1.29), respectively. This interval estimate is quite large on the hazard ratio scale, suggesting that there is not enough information about the treatment difference between the ACEi and the placebo. That is, one cannot even claim that the treatment is no worse than the placebo.

[Figure 1 about here.]

Studies with such low event rates are not uncommon, especially when the primary endpoint is the time to a serious adverse event, for example, in a post-market safety trial. Using an event-driven approach for designing and analyzing an equivalence or non-inferiority trial can be problematic, especially when the observed event rates were lower than their estimated values used at the design stage. On the other hand, using the survival rate differences as the parameters of interest can be quite beneficial for reducing the size and/or the follow-up time of the study. To illustrate this point, in Table 1 we report the point and 0.95 interval estimates for the differences (ACEi minus placebo) of two KM curves at various study time points t's using the above PEACE data. From a clinical point of view, these intervals are tight. The lengths or precisions of these interval estimates depend on the observed number of events, but also depend on the numbers of patients in the risk set at each failure time. For example, at Month 36, the 0.95 confidence interval for the difference of two survival rates is (-0.8, 1.0)%. If we are interested in the average difference of survival functions over the time interval $[0, 60]$ (months), we show in the next section how to construct confidence intervals based on $\hat{D}$. For the present example, the resulting 0.95 confidence interval for this integrated difference is $(-0.7, 0.6)$%. Now, if we are interested in the treatment difference with respect to relatively long term survival for such a patient population, we may consider an integrated difference over a time interval, for example, $[48, 60]$. For this case, the resulting interval is $(-1.4, 1.0)$%. These tight confidence intervals for the differences provide valuable information about the lack of efficacy of the ACEi for these relatively young patients.

[Table 1 about here.]

In the next section, we show how to construct confidence intervals for the integrated difference of two survival functions over a pre-specified time interval. A numerical study was conducted to examine the performance of our proposal under various practical settings. Further remarks about the usage of the integrated difference are given in Section 4.

2. THE DISTRIBUTION OF THE ESTIMATED INTEGRATED DIFFERENCE OF TWO SURVIVAL FUNCTIONS

Let $S_1(\cdot)$ and $S_2(\cdot)$ be the underlying survival functions for $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$, respectively. Let $D(\cdot) = S_2(\cdot) - S_1(\cdot)$. Consider an integrated weighted D_w over the time interval $[t_0, t_1]$:

$$D_w = \bar{w}^{-1} \int_{t_0}^{t_1} w(t)D(t)dt, \tag{2.1}$$

where $w(\cdot)$ is a positive weight function over $[t_0, t_1]$ and $\bar{w} = \int_{t_0}^{t_1} w(t)dt$. Note that if $D(t)$'s are equal to an unknown constant, say, τ, then $D_w = \tau$. If $D(t)$ is not constant over the time interval, one may choose a weight function $w(\cdot)$ such that the resulting D_w is a meaningful summary of the treatment difference over $[t_0, t_1]$. Now, under the usual random sampling setting, let T_{ki} and C_{ki} be the survival and censoring times for the ith subject in the kth treatment group, $k = 1, 2$, $i = 1, \ldots, n_k$. Let $U_{ki} = \min\{T_{ki}, C_{ki}\}$ and $\Delta_{ki} = I\{T_{ki} \leqslant C_{ki}\}$, where $I(\cdot)$ is the indicator function. Furthermore, let $\lambda_k(\cdot)$ be the hazard function of T_{ki}. Let $n = n_1 + n_2$. We assume that $\pi_k = \lim_{n\to\infty} n_k/n > 0$ for $k = 1, 2$.

Now, assume that $Pr(U_{ki} > t_1) > 0, k = 1, 2$. Then, using the uniform consistency property of the KM estimator (Gill, 1983), it is straightforward to show that

$$\hat{D}_w = \bar{w}^{-1} \int_{t_0}^{t_1} w(t)\hat{D}(t)dt, \tag{2.2}$$

consistently estimates D_w. To derive an approximation to the distribution of (2.2), for $t_0 \leqslant t \leqslant t_1$, we first use the following approximation:

$$\hat{S}_k(t) - S_k(t) \approx -S_k(t) \sum_{i=1}^{n_k} \int_0^t \bar{Y}_k^{-1}(u)dM_{ki}(u), \tag{2.3}$$

where $N_{ki}(t) = I\{U_{ki} \leqslant t, \Delta_{ki} = 1\}$, $Y_{ki}(t) = I\{U_{ki} \geqslant t\}$, $\bar{Y}_k(\cdot) = \sum_{i=1}^{n_k} Y_{ki}(\cdot)$, and $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_{ki}(u)\lambda_k(u)du$ (Fleming and Harrington, 1991, p.98). It follows from the martingale central limit theorem (Fleming and Harrington, 1991, ch5) that the right hand side of (2.3) is asymptotically Gaussian over the interval $[t_0, t_1]$.

Next, using a perturbation-resampling method similar to a wild bootstrapping utilized by Lin, Wei and Ying (1993) and Parzen et al. (1997), the distribution of the right hand side of (2.3) can be approximated by the conditional distribution (conditional on the data) of

$$L_k^*(t) = \hat{S}_k(t) \sum_{i=1}^{n_k} Z_{ki} \int_0^t \bar{Y}_k^{-1}(u) dN_{ki}(u), \tag{2.4}$$

where $\{Z_{ki} : k = 1, 2, i = 1, \ldots, n_k\}$ is a random sample from the standard normal, which is independent of the data. It follows that the distribution of $(\hat{D}_w - D_w)$ can be approximated by the conditional distribution of

$$\bar{w}^{-1} \int_{t_0}^{t_1} w(t)[L_2^*(t) - L_1^*(t)]dt. \tag{2.5}$$

Note that the only random quantities in (2.5) are $\{Z_{ki}\}$. In practice, to obtain the distribution of (2.5), we generate a large number M of random samples $\{Z_{ki} : k = 1, 2, i = 1, \ldots, n_k\}$. For each realized normal sample, we compute the corresponding realization of (2.5). Then one may use these M realizations of (2.5) to obtain the sample variance or a robust version thereof as a variance estimate of $\hat{D}_w$. The corresponding confidence intervals for D_w can then be obtained accordingly.

The estimates discussed in the Introduction for the integrated difference of the two survival functions over $[t_0, t_1]$ are obtained with $w(\cdot) = 1$. When $D(t)$'s are equal to a constant τ over $[t_0, t_1]$, we may choose a data dependent weight function to increase the precision in estimating τ. If $\hat{w}(\cdot)$ is uniformly convergent in probability to a deterministic function $w(\cdot)$ over $[t_0, t_1]$, then

$$\hat{D}_{\hat{w}} = \hat{\bar{w}}^{-1} \int_{t_0}^{t_1} \hat{w}(t)\hat{D}(t)dt$$

is a consistent estimator for τ, where $\hat{\bar{w}} = \int_{t_0}^{t_1} \hat{w}(t)dt$. One possible choice of $\hat{w}(t)$ is the reciprocal of a variance estimate of $\hat{D}(t)$ over the above time interval. In the Appendix, we show that under a more general setting, $\hat{D}_{\hat{w}}$ is approximately normal with mean τ. Moreover, the above perturbation-resampling method can still be used to approximate the distribution of $(\hat{D}_{\hat{w}} - \tau)$. It is interesting to note that with the same data set from the PEACE study, the 0.95 confidence intervals for τ using the reciprocal of the variance estimate as the weight are practically identical to or slightly improved over those reported in Table 1 with the constant weight. For example, for the time interval $[t_0, t_1] = [0, 60]$, the confidence interval with the constant weight is (-0.7, 0.6)%. The corresponding empirically weighted one is (-0.4. 0.4)%.

When $D(t)$'s vary over $[t_0, t_1]$, the above simple resampling method or the standard martingale central limit theorem may not be used to approximate the distribution of $(\hat{D}_{\hat{w}} - D_w)$. It is important to note for this general case, one cannot use the results from Pepe and Fleming (1989, 1991) to obtain a large sample approximation to this distribution. In the Appendix, we show how to obtain such an approximation. Note that from the estimation point of view, the empirical weight function $\hat{w}(\cdot)$ should be chosen to have an interpretable summary D_w for the treatment difference over $[t_0, t_1]$.

3. AN EMPIRICAL STUDY

We conducted an extensive numerical study to examine the performance of the new estimation procedure, especially for cases when the event rates are low. For all cases studied, the empirical coverage levels of our interval estimators based on $\hat{D}_w$ were very close to their nominal counterparts even when the crude event rates were only around 3% under various practical settings. For instance, under one of various simulation settings, we mimicked the PEACE study with the aforementioned relatively young patient population. First, for each treatment group, we fitted the observed survival data with a two-parameter Weibull model. We then generated 1000 random samples of survival times via each fitted Weibull model with various sample sizes. Furthermore, we assumed that the censoring distribution for each treatment group is the same as the observed KM estimate. Note that the ranges of both observed KM curves were from 0 to 1. The empirical coverage levels for interval estimates based on $\hat{D}_w$ with a nominal level of 0.95 with various sample sizes and $[t_0, t_1]$ are the entries under the heading “(4.7, 4.5)” for the average crude event rates in Table 2. We also considered cases for which the average crude event rates of the control are about 3%, 8% and 10% by modifying the scale parameters of the above two fitted Weibull models. The empirical coverage levels of 0.95 confidence interval estimates are also reported in Table 2 with various sample sizes and $[t_0, t_1]$. All the entries in the Table are practically equal to their nominal counterparts. Note that for this simulation, we let the weight function $w(\cdot)$ be one.

[Table 2 about here.]

4. REMARKS

In general, the point and interval estimates based on $\hat{D}_w$ for an average difference of two survival curves are easier to interpret than their counterparts for an average ratio of two hazard functions. The choice of $[t_0, t_1]$ for the integrated difference $\hat{D}_w$ of two survival functions depends on the questions for which we would like to have answers from the study. For example, to test equivalence of two groups under a superiority trial setting, Pepe and Fleming (1989, 1991) let $t_0 = 0$ and choose a weight function empirically to increase the power of the test. From the estimation point of view, one may choose a time interval whose members t's are relatively large to examine a *long* survival benefit from the new treatment. Furthermore, one would choose the weight function $w(\cdot)$ with which the resulting D_w is a clinically meaningful summary measure. Although the assumption of proportional hazards or a constant difference of two survival functions is likely violated, the PH estimate or $\hat{D}_w$ provides an average treatment difference over time. On the other hand, when the event rates are low, under an equivalence or non-inferiority study setting, the interval estimate based on $\hat{D}_w$ appears to be a more natural summary metric for quantifying the treatment contrast at the final or interim analysis than its PH counterpart. Therefore, for this case, we strongly recommend using D_w as the primary parameter to design the trial instead of using a conventional, event-driven approach, which may need much more resource to obtain a definite answer to the question regarding the treatment difference.

It is important to note that oftentimes the numbers of events utilized at the design stage for an event driven trial tend to be significantly higher than the observed. This may be due to the improvement of the standard care (or the control), or to publication bias for estimating the historical event rates, or to the investigator's enthusiasm for convincing the sponsor to support the study. The lack of a sufficient number of observed events may cause an early termination of the study. Such conclusion of lack of information can be rather misleading. The fundamental problem of using an estimated hazard ratio as a measure of the treatment contrast is that we do not use the underlying event rate information in designing, monitoring and analyzing the study. On the other hand, the precision of the estimated difference of two KM curves depends on the number of study participants. Using this survival rate difference metric to quantify the treatment difference may lead us to conclude that the two treatments are "equivalent" (not lack of information). Moreover, for the low event rate case with a fixed numbers of study subjects, when the event rates decrease, the precision of the Cox's hazard ratio estimate decreases, but its counterpart of the estimated integrated difference would increase. This interesting feature, coupled with its easy interpretation, makes the integrated difference a better measure for the treatment contrast than its hazard ratio based counterpart.

References

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

Fleming, T. R. and Harrington, D. P. (**1991**), *Counting processes and survival analysis* (New York, N.Y. : Wiley).

Gill, R. D. (**1983**), "Large Sample Behaviour of the Product-Limit Estimator on the Whole Line," The Annals of Statistics **11**, 49–58.

Lin, D. Y., Wei, L. J., and Ying, Z. (**1993**), "Checking the Cox model with cumulative sums of martingale-based residuals," Biometrika **80**, 557–572.

Parzen, M. I., Wei, L. J., and Ying, Z. (**1997**), "Simultaneous Confidence Intervals for the Difference of Two Survival Functions," Scandinavian Journal of Statistics **24**, 309–314.

Pepe, M. S. and Fleming, T. R. (**1989**), "Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data," Biometrics **45**, 497–507.

Pepe, M. S. and Fleming, T. R. (**1991**), "Weighted Kaplan-Meier statistics: Large sample and optimality considerations," Journal of the Royal Statistical Society, Series B **53**, 341–352.

Braunwald E., Domanski M. J., Fowler S. E., 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.

Appendix

In this appendix, we derive the asymptotical properties of the estimator $\hat{D}_{\hat{w}}$. Firstly, we assume that

$$\hat{w}(t) - w(t) = \sum_{k=1}^{2} n_k^{-1} \sum_{i=1}^{n_k} \eta_{ki}(t) + o_p(n^{-\frac{1}{2}}) \tag{A.1}$$

for $t \in [t_0, t_1]$, where $\eta_{k1}(t), \cdots, \eta_{kn_k}(t)$ are n_k independent identically distributed mean zero random processes. Let

$$V_k(t) = \frac{n_k^{\frac{1}{2}} \left[\hat{S}_k(t) - S_k(t)\right]}{S_k(t)}, \quad k = 1, 2.$$

Then $V_k(t)$ is asymptotically equivalent to

$$-n_k^{-\frac{1}{2}} \sum_{i=1}^{n_k} \int_0^t \frac{dM_{ki}(u)}{G_k(u)}$$

where $G_k(u) = Pr(U_{ki} \geqslant u)$ and $M_{ki}(u) = N_{ki}(u) - \int_0^u I(U_{ki} \geqslant s)\lambda_k(s)ds$, $k = 1, 2$ (Fleming and Harrington, 1991, p.98).

Now we have

$$\begin{aligned}
\mathcal{W} &= \left(\frac{n_1 n_2}{n}\right)^{\frac{1}{2}} \left(\hat{D}_w - D_w\right) \\
&= \left(\frac{n_1 n_2}{n}\right)^{\frac{1}{2}} \bar{w}^{-1} \int_{t_0}^{t_1} w(t) \left[\hat{D}(t) - D(t)\right] dt \\
&\quad + \left(\frac{n_1 n_2}{n}\right)^{\frac{1}{2}} \int_{t_0}^{t_1} \left[\frac{\hat{w}(t)}{\hat{\bar{w}}} - \frac{w(t)}{\bar{w}}\right] \hat{D}(t) dt.
\end{aligned}$$

For the first term, applying the integration by parts and Gill (1983, Theorem 2.1), we have

$$\begin{aligned}
& n_k^{\frac{1}{2}} \int_{t_0}^{t_1} w(t) \left[\hat{S}_k(t) - S_k(t)\right] dt \\
=& \int_{t_0}^{t_1} w(t) S_k(t) V_k(t) dt \\
=& \int_0^{t_1} \left[\int_{t \vee t_0}^{t_1} w(u) S_k(u) du\right] dV_k(t) \\
=& -n_k^{-\frac{1}{2}} \sum_{i=1}^{n_k} \int_0^{t_1} \left[\int_{t \vee t_0}^{t_1} w(u) S_k(u) du\right] \frac{dM_{ki}(t)}{G_k(t)},
\end{aligned}$$

where $t \vee t_0 = \max\{t, t_0\}$. For the second term

$$\begin{aligned}
& \left(\frac{n_1 n_2}{n}\right)^{\frac{1}{2}} \int_{t_0}^{t_1} \left[\frac{\hat{w}(t)}{\hat{\bar{w}}} - \frac{w(t)}{\bar{w}}\right] \hat{D}(t) dt \\
=& \ \bar{w}^{-1} \left[\pi_2^{\frac{1}{2}} n_1^{-\frac{1}{2}} \sum_{i=1}^{n_1} \int_{t_0}^{t_1} \eta_{1i}(t) \{D(t) - D_w\} dt + \pi_1^{\frac{1}{2}} n_2^{-\frac{1}{2}} \sum_{i=1}^{n_2} \int_{t_0}^{t_1} \eta_{2i}(t) \{D(t) - D_w\} dt\right] + o_p(1).
\end{aligned}$$

Therefore $\mathcal{W}$ can be written as $n_1^{-\frac{1}{2}} \sum_{i=1}^{n_1} \tau_{1i} + n_2^{-\frac{1}{2}} \sum_{i=1}^{n_2} \tau_{2i}$ where

$$\tau_{1i} = \pi_2^{\frac{1}{2}} \bar{w}^{-1} \left[-\int_0^{t_1} \left\{\int_{t \vee t_0}^{t_1} w(u) S_1(u) du\right\} \frac{dM_{1i}(t)}{G_1(t)} + \int_{t_0}^{t_1} \eta_{1i}(t) \{D(t) - D_w\} dt\right]$$

and

$$\tau_{2i} = \pi_1^{\frac{1}{2}} \bar{w}^{-1} \left[\int_0^{t_1} \left\{\int_{t \vee t_0}^{t_1} w(u) S_2(u) du\right\} \frac{dM_{2i}(t)}{G_2(t)} + \int_{t_0}^{t_1} \eta_{2i}(t) \{D(t) - D_w\} dt\right].$$

By central limit theorem, $\mathcal{W}$ converges weakly to a normal distribution with mean zero and variance $E(\tau_{1i}^2+\tau_{2i}^2)$, which can be estimated by its empirical counterpart, i.e.

$$\frac{n_2}{n_1 n\hat{\bar{w}}^2}\sum_{i=1}^{n_1}\left[-\int_0^{t_1}\left\{\int_{t\vee t_0}^{t_1}\hat{w}(u)\hat{S}_1(u)du\right\}\frac{d\hat{M}_{1i}(t)}{\hat{G}_1(t)}+\int_{t_0}^{t_1}\hat{\eta}_{1i}(t)\{\hat{D}(t)-\hat{D}_w\}dt\right]^2$$

$$+\frac{n_1}{n_2 n\hat{\bar{w}}^2}\sum_{i=1}^{n_1}\left[\int_0^{t_1}\left\{\int_{t\vee t_0}^{t_1}\hat{w}(u)\hat{S}_2(u)du\right\}\frac{d\hat{M}_{2i}(t)}{\hat{G}_2(t)}+\int_{t_0}^{t_1}\hat{\eta}_{2i}(t)\{\hat{D}(t)-\hat{D}_w\}dt\right]^2,$$

where $\hat{G}_k(t)=n_k^{-1}\sum_{i=1}^{n_k}I(U_{ki}\geqslant t)$, $\hat{M}_{ki}(t)=N_{ki}(t)-\int_0^t I(U_{ki}\geqslant u)d\hat{\Lambda}_k(u)$, and

$$\hat{\Lambda}_k(u)=n_k^{-1}\sum_{i=1}^{n_k}\int_0^t\frac{dN_{ki}(u)}{\hat{G}_k(u)},$$

$k=1,2$.

Note that when $D(t)=\tau, t\in[t_0,t_1]$, the second term disappears and the simple resampling method proposed in the paper still provides valid inference for τ as long as $\hat{w}(t)$ is uniformly convergent in probability to a deterministic positive function $w(t)$.

Note that (A.1) is a mild condition. For example, we can show that the weight function

$$\hat{w}(t)=\left\{\sum_{k=1}^2\hat{S}_k(t)^2\int_0^t\frac{d\hat{\Lambda}_k(u)}{\hat{G}_k(u)}du\right\}^{-1},$$

the reciprocal of a variance estimate of $\hat{D}(t)$, satisfies (A.1). Specifically, firstly, it follows from the uniform consistency of $\hat{S}_k(\cdot)$, $\hat{G}_k(u)$ and $\hat{\Lambda}_k(u)$, $\hat{w}(t)$ is a uniform consistent estimator for

$$w(t)=\left\{\sum_{k=1}^2 S_k(t)^2\int_0^t\frac{d\Lambda_k(u)}{G_k(u)}du\right\}^{-1}.$$

Furthermore, with routine algebraic operations, we can show that

$$\hat{w}(t)-w(t)=\sum_{k=1}^2 n_k^{-1}\sum_{i=1}^{n_k}\eta_{ki}(t)+o_p(n^{-\frac{1}{2}}),$$

where

$$\eta_{ki}(t)=w(t)^2S_k(t)^2\int_0^t\frac{dM_{ki}(u)}{G_k^2(u)}\left[2\left(\int_0^t\frac{d\Lambda_k(v)}{G_k(v)}dv\right)G_k(u)+\{I(U_{ki}\geqslant u)-G_k(u)\}-1\right].$$

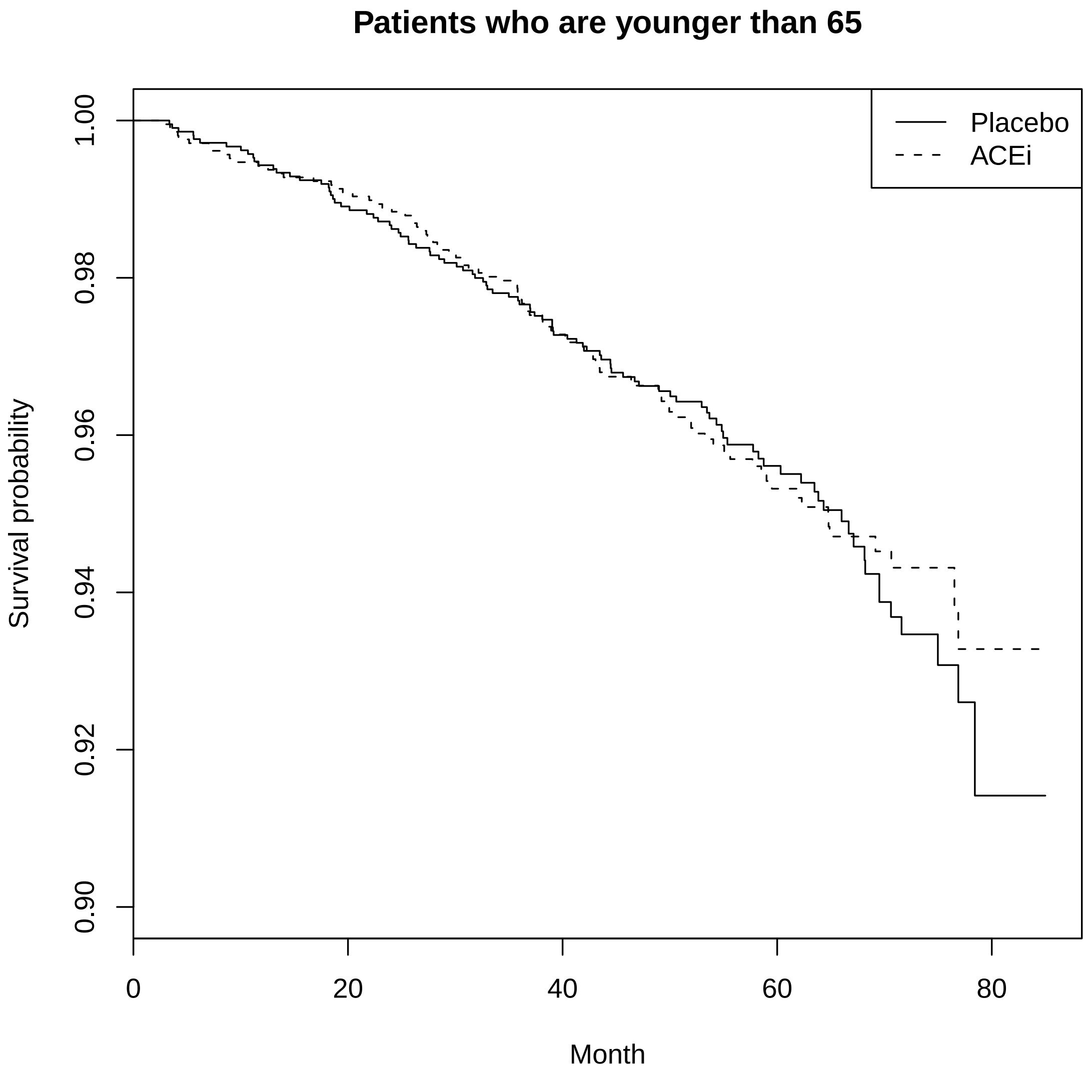

Figure 1. The Kaplan-Meier estimates for the survival functions of patients who are younger than 65 in the PEACE study

Table 1

Point estimate (PE) and 0.95 confidence interval (CI) for various treatment contrast measures (ACEi vs. placebo) for patients who are younger than 65 in the PEACE study

	PE	CI	Length
Survival rate difference (%)			
Month			
24	0.2	(-0.4, 0.9)	1.3
36	0.1	(-0.8, 1.0)	1.8
48	0.0	(-1.1, 1.1)	2.2
60	-0.3	(-1.7, 1.1)	2.8
Integrated difference (%)			
Time interval			
$[0, 60]$	-0.0	(-0.7, 0.6)	1.3
$[36, 60]$	-0.1	(-1.2, 0.9)	2.1
$[48, 60]$	-0.2	(-1.4, 1.0)	2.4
Hazard ratio			
	0.97	(0.73, 1.29)	0.55

Table 2
Empirical coverage levels of 0.95 confidence intervals for the integrated difference of two Weibull survival functions over interval $[t_0, t_1]$

Sample size for each group	Average crude event rates (%) (placebo, ACEi)			
	(3.0, 3.0)	(4.7, 4.5)[1]	(8.0, 7.6)	(10.0, 9.5)
	$[t_0, t_1] = [0, 60]$			
500	95.9	95.8	94.6	94.4
1000	94.6	94.3	95.1	95.6
2000	94.8	94.8	93.9	95.4
	$[t_0, t_1] = [36, 60]$			
500	96.2	95.9	95.2	94.2
1000	94.1	95.1	95.4	95.1
2000	95.7	94.6	93.3	94.5
	$[t_0, t_1] = [48, 60]$			
500	96.1	94.9	95.1	94.2
1000	94.5	95.0	94.8	95.3
2000	95.6	94.2	94.8	94.6

[1] Based on fitted Weibull models with the data from patients who are younger than 65 in the PEACE study