Harvard University

Harvard University Biostatistics Working Paper Series

Year 2017 *Paper* 212

Quantifying the totality of treatment effect with multiple event-time observations in the presence of a terminal event from a comparative clinical study

Brian Claggett[*] Lu Tian[†] Haoda Fu[‡]

Scott D. Solomon[**] L. J. Wei[††]

[*] Harvard Medical School, bclaggett@partners.org

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

[‡] Eli Lilly, haoda-fu@lilly.com

[**] Harvard Medical School, ssolomon@bwh.harvard.edu

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

Quantifying the totality of treatment effect with multiple event-time observations in the presence of a terminal event from a comparative clinical study

Brian Claggett, Lu Tian, Haoda Fu, Scott D. Solomon, and L. J. Wei

Abstract

To evaluate the totality of one treatment's benefit/risk profile relative to an alternative treatment via a longitudinal comparative clinical study, the timing and occurrence of multiple clinical events are typically collected during the patient's followup. These multiple observations reflect the patient's disease progression/burden over time. The standard practice is to create a composite endpoint from the multiple outcomes, the timing of the occurrence of the first clinical event, to evaluate the treatment via the standard survival analysis techniques. By ignoring all events after the composite outcome, this type of assessment may not be ideal. Various parametric or semi-parametric procedures have been extensively discussed in the literature for the purposes of analyzing multiple event-time data. Many existing methods were developed based on extensive model assumptions. When the model assumptions are not plausible, the resulting inferences for the treatment effect may be misleading. In this article, we propose a simple, non-parametric inference procedure to quantify the treatment effect which has an intuitive, clinically meaningful interpretation. We use the data from a cardiovascular clinical trial for heart failure to illustrate the procedure. A simulation study is also conducted to evaluate the performance of the new proposal.

Quantifying the totality of treatment effect with multiple event-time observations in the presence of a terminal event from a comparative clinical study

BRIAN CLAGGETT[1], LU TIAN[2], HAODA FU[3],
SCOTT D SOLOMON[4], and LEE-JEN WEI[5]

[1]Division of Cardiovascular Medicine, Harvard Medical School,
bclaggett@partners.org.

[2]Department of Biomedical Data Science, Stanford University,
lutian@stanford.edu

[3]Eli Lilly and Company.
haoda-fu@lilly.com

[4]Division of Cardiovascular Medicine, Harvard Medical School,
ssolomon@bwh.harvard.edu

[5]Department of Biostatistics, Harvard University.
wei@sdac.harvard.edu

January 9, 2017

Abstract

To evaluate the totality of one treatment's benefit/risk profile relative to an alternative treatment via a longitudinal comparative clinical study, the timing and occurrence of multiple clinical events are typically collected during the patient's followup. These multiple observations reflect the patient's disease progression/burden over time. The standard practice is to create a composite endpoint from the multiple outcomes, the timing of the occurrence of the first clinical event, to evaluate the treatment via the standard survival analysis techniques. By ignoring all events after the composite outcome, this type of assessment may not be ideal. Various parametric or semi-parametric procedures have been extensively discussed in the literature for the purposes of analyzing multiple event-time data. Many existing methods were developed based on extensive model assumptions. When the model assumptions are not plausible, the resulting inferences for the treatment effect may be misleading. In this article, we propose a simple, non-parametric inference procedure to quantify the treatment effect which has an intuitive, clinically meaningful interpretation. We use the data from a cardiovascular clinical trial for heart failure to illustrate the procedure. A simulation study is also conducted to evaluate the performance of the new proposal. Clinical Trials; Composite Endpoint; Multiple Outcomes; Non-Parametric Estimation; Counting process; Survival analysis; Wei-Lin-Weissfeld procedure.

1 Introduction

In a longitudinal clinical study, each patient may experience any of several clinical events at various time points during the follow-up period. Such multiple event-time observations provide a temporal profile of the patients disease burden or progression.

An important question is how to utilize these observations collectively, for instance, to evaluate a new therapy vs. the standard care from a risk-benefit perspective. A common practice is to consider either the time from enrollment or randomization to a specific event or to the first occurrence of one of a collection of pre-specified clinical events as the study's primary endpoint and to then analyze such data using standard inference procedures from survival analysis. Such approaches, however, may not utilize all relevant information to fully answer the clinical question of interest.

As an example, a randomized, comparative clinical trial, "Beta-Blocker Evaluation of Survival Trial (BEST)," was conducted to evaluate whether the beta-blocking drug, bucindolol, would benefit patients with advanced chronic heart failure (BEST, 2001). For this study, there were 2708 patients enrolled, randomized to receive either placebo or the beta-blocker, who were then followed for an average of two years. The patients overall survival time was chosen as the primary endpoint of the study. For the comparison of the two treatment groups, the p-value of the two-sample logrank test was 0.11 with the 0.95 confidence interval for the hazard ratio of (0.78, 1.02), numerically, but not significantly, in favor of the beta-blocker. Although mortality is an important endpoint, an evaluation of the beta-blockers benefits and risks should also include morbidity for chronic heart failure patients over the course of the study. Clinically important morbidity events for these patients are, for instance, hospitalization for worsening heart failure (WHF), non-heart failure hospitalization (NHFH), myocardial infarction (MI), and heart transplant (HT). The BEST study is a typical cardiovascular trial for which the times to non-fatal events prior to a terminal event (for example, death) can be potentially observed for each patient. If we follow the conventional approach using a composite endpoint, that is, the time of the first occurrence of any of the above five distinct events as the endpoint, the resulting

Kalpan-Meier curves for two arms are given in Figure 1. The 0.95 confidence interval for the hazard ratio is (0.85, 1.02) and the p-value of the logrank test is 0.10. Furthermore, if we consider the distribution of each specific component event, it is apparent that the composite event is more often an occurrence of non-heart failure hospitalization and less often worsening heart failure in the bucindolol arm (Table 1), even though each of these types of events occurs in fewer patients randomized to bucindolol than in patients randomized to placebo. Like the results from the mortality analysis, the beta-blocker has only modest statistical evidence of benefit in this population with respect to this composite outcome.

Table 1: Total number of patients experiencing each type of event, and specific type of clinical events represented by the composite outcome, by treatment group

Event type	Placebo	Bucindolol
Worsening HF	569 (42%)	476 (35%)
Non-HF Hosp	634 (47%)	619 (46%)
Death	449 (33%)	411 (30%)
MI	85 (6%)	46 (3%)
Transplant	41 (3%)	29 (2%)
Composite	971 (72%)	931 (69%)
Composite event type		
Worsening HF	393 (40%)	341 (37%)
Non-HF Hosp	445 (46%)	466 (50%)
Death	99 (10%)	103 (11%)
MI	32 (3%)	18 (2%)
Transplant	2 (<1%)	3 (<1%)
Total	971	931

If the clinical questions regarding the risks and benefits of bucindolol extend beyond the simple analysis of mortality or of the occurrence of the first composite event, several novel statistical procedures for comparing two groups may be used to analyze such multiple event time observations (Wei et al., 1989; Li and Lagakos, 1998; Wang

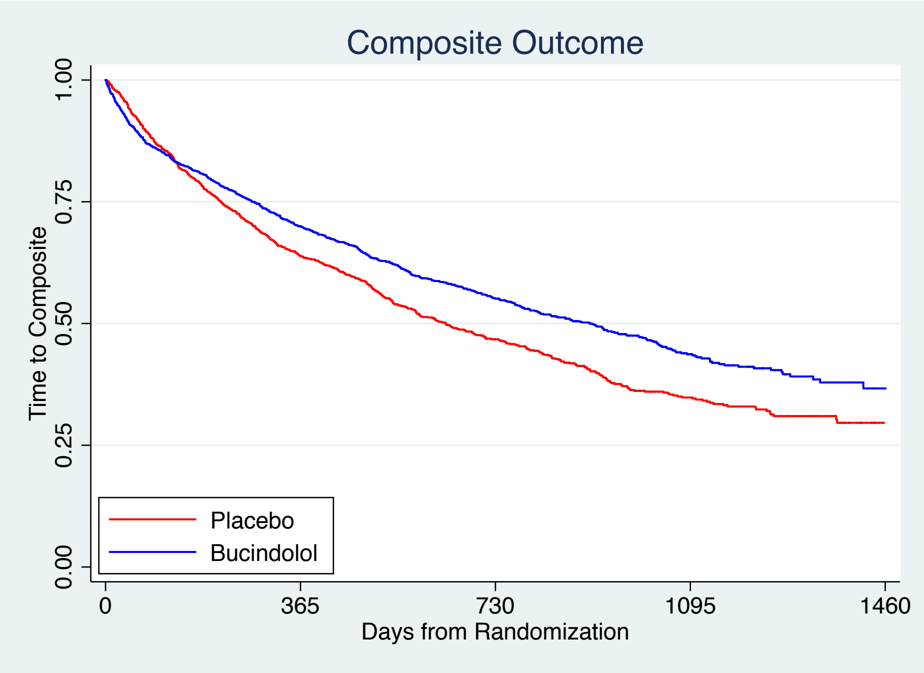

Figure 1: Time to first occurrence of worsening heart failure, non-HF hospitalization, heart transplant, myocardial infarction, or death.

et al., 2001; Wang and Chiang, 2002; Ghosh and Lin, 2003; Lin et al., 2000; Huang and Wang, 2011; Wang and Huang, 2014). These methods generally utilize model-based parameters to quantify the between-group difference. As is the case with univariate survival analysis, when the model assumptions are not plausible, the resulting estimates for the parameters may be difficult to interpret clinically (Kalbfleisch and Prentice, 1981; Struthers and Kalbfleisch, 1986; Lin and Wei, 1989; Hernán, 2010; Uno et al., 2014, 2015).

In this article, in order to include both mortality and morbidity events beyond the first composite endpoint, we consider the patient's endpoint based on a reverse counting process, $R(t)$ over time t, which provides the profile of the multiple event times that comprise the composite outcome above. For example, with the aforementioned five clinical events: WHF, NHFH, MI, HT, and death, in the BEST study, Figure 2 shows several realizations of the $R(\cdot)$ process. Each realization is a downward step function starting with a y-axis value of 5, the number of distinct types of event under consideration. At the time of an occurrence of a non-terminal (non-fatal) event, $R(\cdot)$ drops by one unit, but at the time of the terminal event, $R(\cdot)$ drops to zero. At a specific time t, $R(t)$ represents the number of the composite events not experienced at t. The area under this step function at t, $A(t)$, is the sum of five event-free survival times up to t. For example, for the first realization of $R(\cdot)$ in Figure 2, the observed $A(48)$ is 118 (months). That is, this patient enjoyed 10 months of HF-free survival, 18 months free of MI-free survival, 30 months of HT-free survival, 30 months of NHFH-free survival, and 30 months of overall survival. The cumulative total of these is 118 months of event-free survival. Noting that the ideal case of a patient without any increase in disease burden over the study period of interest (i.e. here, 48 months) would correspond to $A(48) = 240(= 48 \times 5)$ months, this particular patient experienced $49\%(= 118/240)$ of the maximum possible cumulative event-free survival, or conversely, $51\%(= 1 - 118/240)$ of the maximum possible disease burden over this time period, as measured by this combination of morbidity and mortality. The values $A(t)$ or the above ratio, $P(t)$, for example, would be clinically meaningful summaries for the temporal profile of patient health with regards to these multiple event times up to time t. Note that for the second realization in Figure 2, only one of the five outcomes is observed prior to the patient's censoring at month 30. For this patient, $A(48)$ is not fully observed, but the available partial information indicates that $A(48) \in (140, 212]$ months (140 if the patient died the day after censoring; 212 if the patient experienced no subsequent events until month 48) and $P(48)$ is between

0.12 and 0.42. It is important to note that in the presence of a terminal event such as death, the standard forward counting process as the patient's endpoint is problematic, since this process is not well defined after death.

For the comparison of two groups, the difference or ratio of the two expected values $E(R(t))$, $E(A(t))$, or $E(P(t))$ is a clinically interpretable, model-free summary to quantify the between-group contrast. In this paper, we present inference procedures for handling one- and two-sample problems. All of the proposals are illustrated with the data from the BEST study. Note that for the case with a single event time observation for each patient, $E(R(t))$ reduces to the survival rate $S(t)$ and $E(A(t))$ is the so-called restricted mean survival time at time t, which has been extensively studied, for example, by Karrison (1987); Zucker (1998); Royston and Parmar (2011); Zhao et al. (2012); Tian et al. (2014); Uno et al. (2014); Trinquart et al. (2016); A'Hern (2016). Furthermore, classical methods such as Andersen and Gill (1982); Fine and Gray (1999); Lin et al. (2000) either treat the terminal events as censoring, consider the patient to be at risk for non-fatal events even after death, or extrapolate the counting process of nonfatal and terminal events by assuming that there is no nonfatal event after death. The validity of the former approach relies on the non-informative censoring assumption, which unrealistically assumes that the terminal event is independent of the non-fatal events during the follow-up. The latter does not differentiate nonfatal from terminal events and may yield misleading comparisons when the mortality rate is very different between two groups. For example, the low incidence rate of one arm may reflect higher mortality rate rather than a real clinical benefit. Other methods such as Liu et al. (2004); Rondeau et al. (2007) explicitly model the joint distribution of non-fatal and terminal events and produce estimators of the treatment effect on non-fatal and fatal events separately. These methods are heavily model-dependent and it is difficult to combine the two estimates into a single summary in a setting of binary decision making. Mao and Lin (2016) recently proposed a semiparametric model for a composite outcome based on pre-specified weights for different types of events, relying on an assumption of multiplicative effects on the marginal rate function.

2 One- and two-sample inference procedures

Suppose that for each study subject, there are $(K+1)$ distinct types of events of interest, which can be potentially observed during the study follow-up. Also assume that the $(K+1)$th event is the only terminal event. Let $R(\cdot)$ be the reverse counting process described in the Introduction with respect to these $K+1$ events. In this Section, we are interested in making inferences about the parameters $E(R(\cdot))$, $E(A(\cdot))$, and $E(P(\cdot))$. Now, let $T_k, k = 1, \cdots, K+1$, be the minimum of $\tilde{T}_k$ and $\tilde{T}_{K+1}$, where $\tilde{T}_k$ is the time to the first occurrence of the kth type of event. Then,

$$R(t) = \sum_{k=1}^{K+1} I(T_k \geq t), \tag{2.1}$$

where $I(\cdot)$ is the indicator function,

$$A(t) = \sum_{k=1}^{K+1} A_k(t), \tag{2.2}$$

where $A_k(t)$ is the minimum of T_k and t, and

$$P(t) = 1 - \frac{A(t)}{t(K+1)}. \tag{2.3}$$

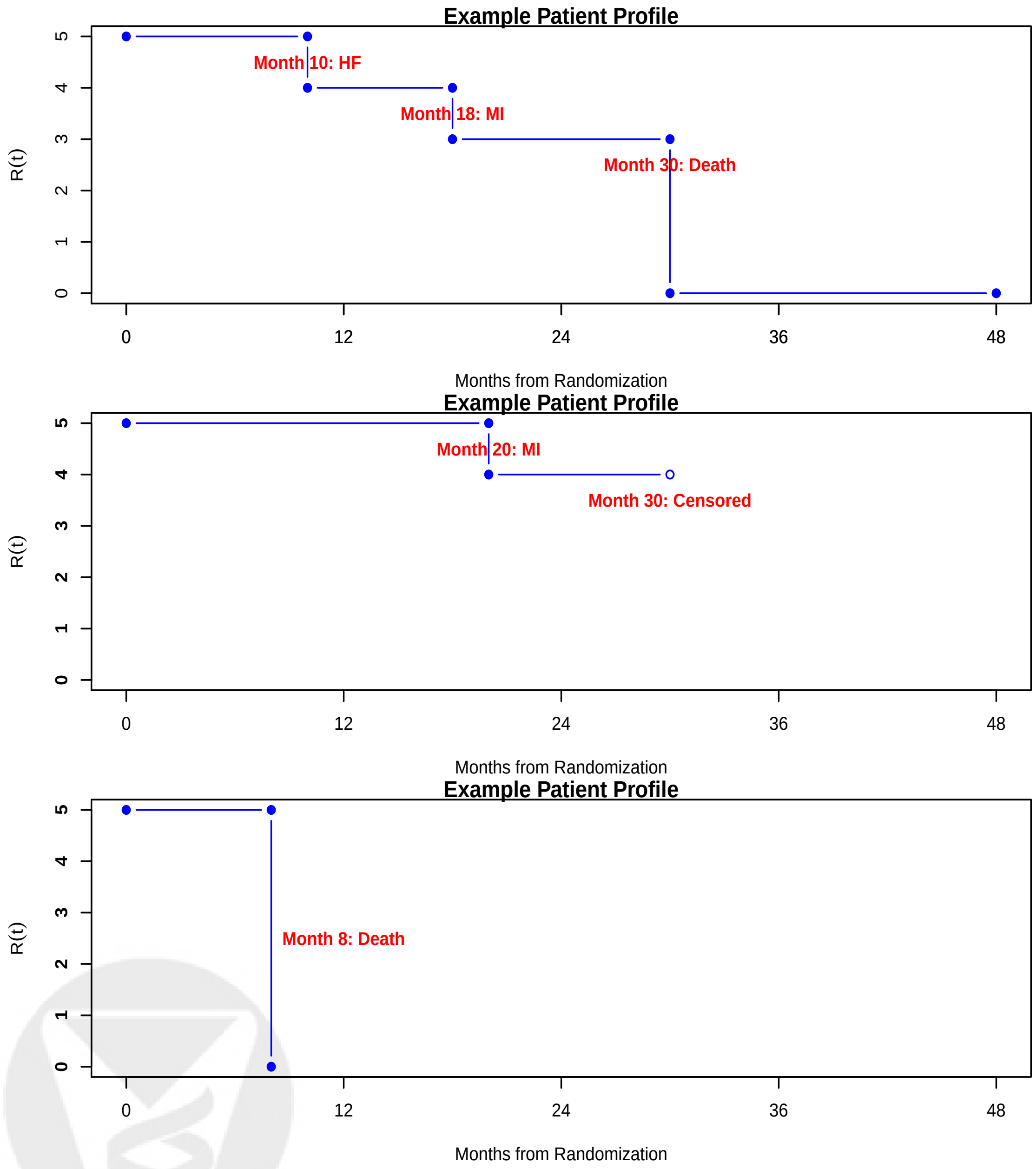

Figure 2: Profile of observed data from three hypothetical patients from randomization to end of follow-up.

Note that $E(A_k(t))$ is the restricted mean survival time up to time t for T_k, which is the area under the survival curve for T_k up to time t.

The above processes may not be observed completely if the terminal event time $\tilde{T}_{K+1}$ is censored by a random variable C, which is assumed to be independent of $\tilde{T}_{K+1}$, as well as the non-terminal event times $\tilde{T}_1, \ldots, \tilde{T}_K$. Let X_k be the minimum of T_k and C, $\Delta_k = 1$ if T_k is observed and zero, otherwise, for $k = 1, \cdots, K$, and $\bar{\Delta} = 1$ if $\tilde{T}_{K+1}$ is observed and zero, otherwise. The data, $(\{R_i(t), 0 \le t \le C_i(1 - \bar{\Delta}_i) + \tau^* \bar{\Delta}_i\}, \bar{\Delta}_i)$, $i = 1, \cdots, n$, where τ^* is the maximum study followup time, consist of n independent copies of $(\{R(t), 0 \le t \le C(1 - \bar{\Delta}) + \tau^* \bar{\Delta}\}, \bar{\Delta})$.

Using (2.1-2.3), $E(R(t))$, $E(P(t))$, and $E(A(t))$ can be consistently estimated with these n sets of possibly incomplete observations by

$$\widehat{E(R)}(t) = \sum_{k=1}^{K+1} \hat{S}_k(t), \tag{2.4}$$

where $\hat{S}_k(\cdot)$ is the Kaplan-Meier (KM) estimate for T_k based on $\{X_{ik}, \Delta_{ik}, i = 1, \cdots, n\}$ for $k = 1, \cdots, K + 1$; and

$$\widehat{E(A)}(t) = \sum_{k=1}^{K+1} \widehat{E(A_k)}(t), \tag{2.5}$$

where $\widehat{E(A_k)}(t)$ is the area under the KM curve $\hat{S}_k(\cdot)$ up to time t.

It is important to note that although the censoring variable C is assumed to be independent of all the event times, the outcome processes $R(t)$, $A(t)$, and $P(t)$ may be correlated with C. Such an induced dependence results in some technical difficulty for deriving the large sample properties of $\widehat{E(R)}(\cdot)$ and $\widehat{E(A)}(\cdot)$ (Glasziou et al., 1990; Lin, 2003). For the present case, due to the decompositions (2.4) and (2.5), one may use similar techniques, for example, as in Wei et al. (1989); Li and Lagakos (1998). to justify the large sample mean-zero Gaussian approximations to the distributions of $\{\widehat{E(R)}(t) - E(R)(t)\}$ and $\{\widehat{E(A)}(t) - E(A)(t)\}$ as processes over time, such that $t \leq \tau$, $pr(X_k > \tau) > 0$ for all k . In practice, approximations to these distributions can be obtained via a perturbation-resampling method. Specifically, a perturbed version of each KM estimate is

$$S_k^*(t) = \exp\left[-\sum_{i=1}^{n} \int_0^t \frac{V_i d\{I(u \leq T_{ik})\Delta_{ik}\}}{\sum_{l=1}^{n} V_l I(X_{lk} \geq u)}\right] \tag{2.6}$$

where $t \leq \tau$, and $\{V_i : i = 1, \ldots n\}$ is a random sample of size n from the standard exponential distribution. For each realization of random weights $\{V_i\}$, let $\widehat{E(R^*)}(t) = \sum_k S_k^*(t)$, and $\widehat{E(A^*)}(t) = \sum_k \widehat{E(A^*)}_k(t)$, where $\widehat{E(A^*)}_k(t)$ is the area under the KM curve $S_k^*(\cdot)$ up to time t (Royston and Parmar, 2011; Zhao et al., 2012; Tian et al., 2014). Then the distribution of $\sqrt{n}\{\widehat{E(R)(\cdot)} - E(R)(\cdot)\}$ can be approximated by the distribution of $\sqrt{n}\{\widehat{E(R^*)(\cdot)} - \widehat{E(R)(\cdot)}\}$ with a large number, M, of realizations of random weights $\{V_i\}$. Denote the observed variance as $\hat{\sigma}_R^2(\cdot)$. Similarly, the distribution of $\sqrt{n}\{\widehat{E(A)(\cdot)} - E(A)(\cdot)\}$ can be approximated by the distribution of $\sqrt{n}\{\widehat{E(A^*)(\cdot)} - \widehat{E(A)(\cdot)}\}$ with the corresponding variance estimate $\hat{\sigma}_A^2(\cdot)$. Thus, a $(1 - \alpha)$ confidence interval for $E(R)(\cdot)$, for $t \leq \tau$, is given by

$$\left(\widehat{E(R)(\cdot)} - z_{1-\alpha/2} n^{-1/2} \hat{\sigma}_R(\cdot), \widehat{E(R)(\cdot)} + z_{1-\alpha/2} n^{-1/2} \hat{\sigma}_R(\cdot)\right),$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)th$ quantile of the standard normal distribution. To preserve the range of $E(R)(\cdot) \in [0, K + 1]$, we may also first construct a confidence interval of $g^{-1}(E(R)(\cdot))$ based on the proposed perturbation method and then transform the resulting confidence interval using $g(\cdot)$ to obtain an appropriate confidence interval for $E(R)(\cdot)$, where $g(\cdot)$ is a given monotone function $(-\infty, +\infty) \to [0, K+1]$. Similarly, a $(1-\alpha)$ confidence interval for $E(A)(\cdot)$ is given by

$$\left(\widehat{E(A)(\cdot)} - z_{1-\alpha/2} n^{-1/2} \hat{\sigma}_A(\cdot), \widehat{E(A)(\cdot)} + z_{1-\alpha/2} n^{-1/2} \hat{\sigma}_A(\cdot)\right).$$

This resampling technique has been utilized in dealing with various survival analysis problems (Park and Wei, 2003; Cai et al., 2010).

Now, suppose we are interested in constructing a simultaneous confidence band for $E(R)(\cdot)$ or $E(A)(\cdot)$ over a specific range $t \in [a, b]$, where a is larger than the first observed event time and b is smaller than the largest observed follow-up time. The equal precision $(1-\alpha)$ confidence bands (Gilbert et al., 2002) can be constructed by

$$\left(\widehat{E(R)(\cdot)} - c_\alpha n^{-1/2} \hat{\sigma}_R(\cdot), \widehat{E(R)(\cdot)} + c_\alpha n^{-1/2} \hat{\sigma}_R(\cdot)\right)$$

and

$$\left(\widehat{E(A)(\cdot)} - d_\alpha n^{-1/2} \hat{\sigma}_A(\cdot), \widehat{E(A)(\cdot)} + d_\alpha n^{-1/2} \hat{\sigma}_A(\cdot)\right),$$

where c_α is chosen such that

$$pr\left(\sup_{t \in [a,b]} \left| \frac{\widehat{E(R^*)}(t) - \widehat{E(R)}(t)}{\hat{\sigma}_R(t)} \right| > c_\alpha \right) = \alpha,$$

and d_α is chosen such that

$$pr\left(\sup_{t \in [a,b]} \left| \frac{\widehat{E(A^*)}(t) - \widehat{E(A)}(t)}{\hat{\sigma}_A(t)} \right| > d_\alpha \right) = \alpha.$$

With the data from the placebo arm of the BEST study for the five distinct events discussed in the Introduction, Figure 3 gives the estimate $\widehat{E(R)}(t)$ with the

0.95 pointwise confidence intervals and simultaneous confidence bands for $E(R)(t)$ for $1 \leq t \leq 48$ months based on $M = 500$ sets of perturbed data. These bands are quite informative; for example, in the placebo group, at t = 48 months, on average, there are 2.00 events not occurring before the death with a 0.95 pointwise confidence interval of (1.81, 2.18). The estimated sum of all the event-free survival times, $\widehat{E(A)}(48)$, is 150.8 months with a 0.95 confidence interval of (146.8, 154.8) months. Correspondingly, the estimated proportion of maximum morbidity/mortality experienced was $\widehat{E(P)}(48)$ is 0.372 (0.355, 0.388).

Now, if we are interested in making inferences about the difference $\mathbb{D}_R(\cdot)$ of $E(R_j)(\cdot)$ between two treatment groups $j(=0,1)$, the resulting $\hat{\mathbb{D}}_R(\cdot) = \widehat{E(R_1)}(\cdot) - \widehat{E(R_0)}(\cdot)$ can be obtained via the corresponding empirical counterparts, $\widehat{E(R_j)}(\cdot)$. The distribution of $\hat{\mathbb{D}}_R(\cdot)$ can be approximated via the aforementioned resampling method. Our procedure is an extension of the proposal by Parzen et al. (1997) for the case with the univariate event time observations. The difference $\mathbb{D}_A(t)$ of two $E(A)(t)$'s can be estimated by its counterpart via $\hat{\mathbb{D}}_A(t) = \widehat{E(A_1)}(t) - \widehat{E(A_0)}(t)$. With the data from the BEST study, Figure 4 shows the estimated $\hat{E}(R)(\cdot)$ process for both bucindolol and the placebo groups, along with the corresponding contrast $\hat{\mathbb{D}}_R(\cdot)$ between the beta-blocker and the control arms. At $t = 24$ months, the estimated difference is 0.19 events with a 0.95 confidence interval of (0.03, 0.36). At $t = 48$ months, the estimated difference is 0.18 but with a wider 0.95 confidence interval of (-0.09, 0.46). Note that for each of these comparisons, no information is used regarding the temporal profile of events occurring prior to the selected time point. In order to utilize both the occurrence and the timing of the events, we may used the estimated cumulative difference in total event-free survival time $\mathbb{D}_A(t)$. At the end of followup, this is 7.6 months with a 0.95 confidence interval of (1.5, 13.7) months, demonstrating a significant overall beneficial effect of the active therapy over placebo. Alternatively, this overall treatment difference can be expressed as $\hat{\mathbb{R}}_A(t) = \widehat{E(A_1)}(t)/\widehat{E(A_0)}(t) = 1.05(1.01, 1.09)$, indicating an estimated 5% increase in event-free survival time, with $p = 0.015$ for the test of equality between treatment groups. Another interesting expression is via the comparison of the proportion of follow-up time lost to morbidity and mortality, $P_j(t)$. The ratio of these two estimates $\hat{\mathbb{R}}_P(t) = \widehat{E(P_1)}(t)/\widehat{E(P_0)}(t) = 0.92$ (0.85, 0.98), an 8% decrease in morbidity/mortality.

3 Simulations

In order to assess the properties of the proposed area under the curve, $\widehat{E(A)}(t)$, for the purpose of comparing two treatment groups, we performed an extensive simulation, intended to mimic a trial setting similar to that of the BEST trial. In the simulations below, we consider a trial with $N = 1500$ patients followed for a maximum time τ of 4 years, in which there are a total of four clinical event of interest: three non-fatal events in addition to all-cause mortality. In all scenarios, the event times in the placebo group are drawn from Weibull distributions with shape parameter 0.8, and scale parameters $2000, 3000,$ and 4000 for the non-fatal event and 8000 for the fatal events, which correspond to survival probabilities of 46%, 57%, 64%, and 77%, respectively, at the end of the follow-up period. In order to reflect the common scenario in which event times are correlated within patients, we induce a shared frailty parameter drawn from a gamma distribution with an unit mean and variance = 2 (Liu et al., 2004; Rondeau et al., 2007, 2013). In Scenario 0, the treatment has no effect on any of the four clinical outcomes, representing the null hypothesis. We then consider treatment effects which reduce time lost to morbidity/mortality by either 10% (moderate effect)

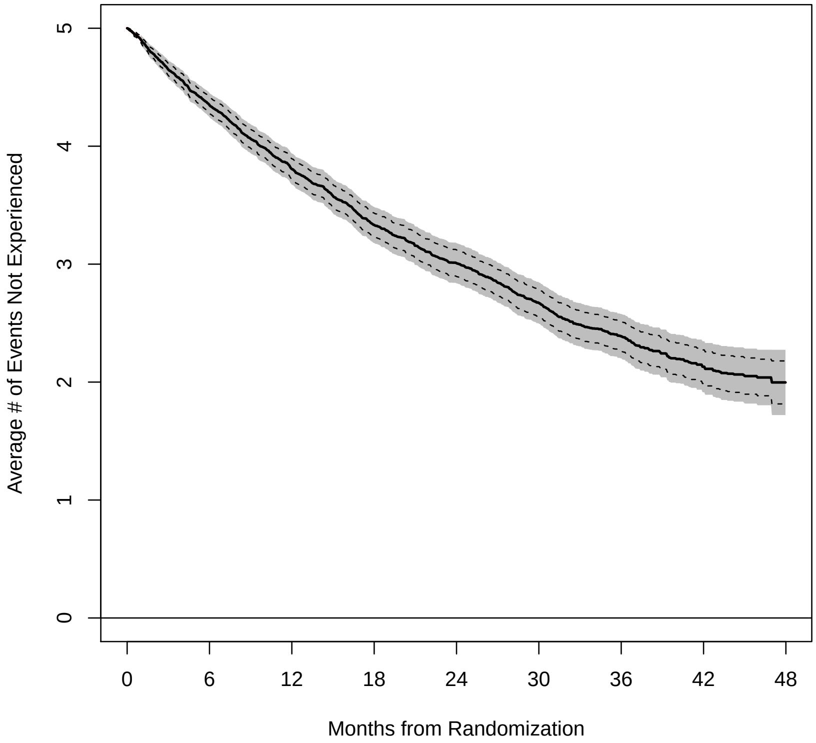

Figure 3: Point and interval estimates of $E(R)(t)$ over time from the placebo arm of the BEST trial. Solid curve represents point estimates, with 0.95 pointwise and simultaneous confidence intervals denoted by dashed lines and gray shading, respectively.

or 20% (strong effect). In Scenario 1, the treatment effect is strong with respect to the two more frequent non-fatal events, but moderate for the other two events. In Scenario 2, the treatment effect is strong with respect to the two less frequent events, but moderate for the more frequent events. In Scenario 3, the treatment effect is strong with respect to all 4 events. Within each scenario, we considered that the treatment effect may manifest itself through a constant reduction in hazard, (i.e.

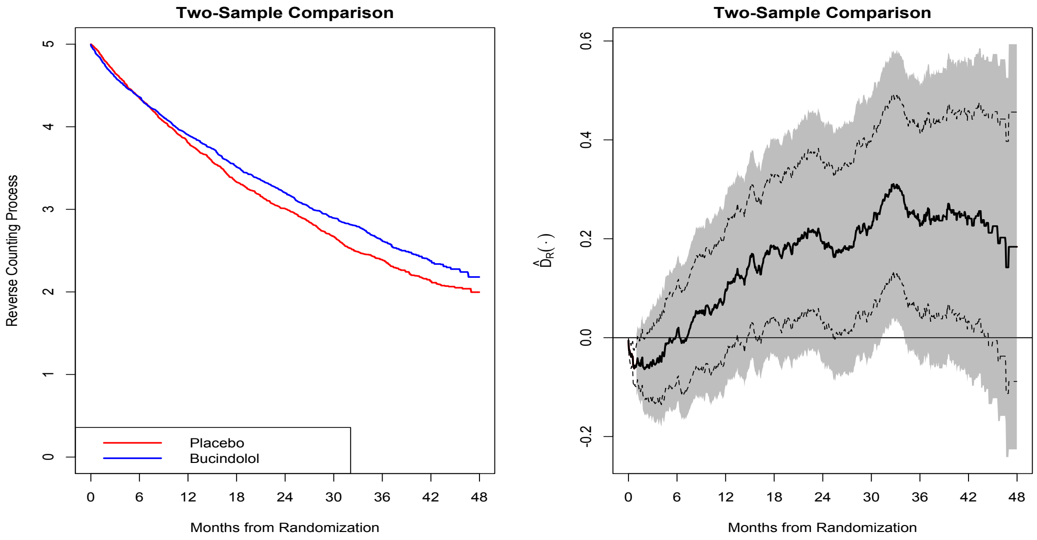

Figure 4: Left: Estimated number of events not yet experienced in each treatment arm. Right: Treatment effect $\hat{\mathbb{D}}_R(\cdot)$ as a function of follow-up time. Solid curve represents point estimates, with 0.95 pointwise and intervals denoted by dashed lines.

the shape parameter remains constant, and the scale parameters is increased in the treated arm, PH assumption), or alternatively, through a delay in event times, such that the treatment and control groups' survival curves become equal at the end of the study, but the treatment group's survival curve is uniformly above the control group's for the duration of the study (i.e. the shape parameter is increased in the treated arm, non-PH assumption). We assume independent administrative censoring, reflecting a hypothetical 5-year trial with 3 years of uniform enrollment, so that every patient is followed for at least two years. We compare the proposed method to the traditional "time-to-first" composite outcome compared via the log-rank test. The Table below shows the proportion of simulated data sets in which the null hypothesis of no treatment effect is rejected at the $\alpha = 0.05$ level.

In Scenario 0, we find that type-I error is well controlled. In Scenarios 1-3, we see

Table 2: Two-Sample Power

Scenario	Treatment Effect Frequent Events	Treatment Effect Less Frequent Events	Proportional Hazards	Method 1: RCP	Method 2: LR (first event)
0	none	none	Yes	5%	5%
1	strong	moderate	Yes	34%	34%
			No	45%	44%
2	moderate	strong	Yes	48%	22%
			No	54%	28%
3	strong	strong	Yes	70%	56%
			No	76%	63%

that the proposed metric has equal or greater power than the standard "time-to-first" event approach in all settings, particularly when the treatment effect is strong with respect to the fatal events and when the PH assumption does not hold.

4 Discussion

Although many statistical methods are currently available to compare two treatment groups in the presence of multiple outcomes, a method that is not dependent on a particular parametric modeling assumption is preferable. The ability to produce estimates of treatment effects which cannot be undermined by model misspecification should be seen as a benefit to investigators, sponsors, and regulators, each of whom rely on the robustness of the inferences drawn from clinical studies. Moreover, an intuitive and interpretable measure of the magnitude of treatment effect expressed in concrete terms such the numbers of days spent event-free or the number of events prevented is quite attractive. For example, the constant intensity or rate function model for recurrent event times (Andersen and Gill, 1982; Lin et al., 2000) may be theoretically interesting, but the results are difficult to interpret, especially when the model assumption is violated.

The methods proposed in this article represent extensions of relatively standard concepts in the analysis of survival data to address to an important open question in the general community of clinical trialists. We note that under certain circumstances, it may be desirable to modify the starting value of the reverse counting process or the relative values of the individual events in the reverse counting process $R(0)$. For example, reducing the starting value to $R(0) = 1$ results in a conventional "time-to-first-event" analysis. One may also desire to implement weights w_k associated with each of the $K + 1$ event types, similar to ad-hoc procedures which have appeared in the clinical literature (Armstrong et al., 2011).

Acknowledgements

This manuscript was prepared using BEST Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the BEST investigators or the NHLBI. This research was partially supported by US NIH grants and contracts.

Supplementary Materials

Technical appendices and additional results are provided in on-line Supplementary Materials.

References

A'Hern, R. P. (2016). Restricted mean survival time: An obligatory end point for time-to-event analysis in cancer trials? *Journal of Clinical Oncology*, page JCO678045.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 10(4):1100–1120.

Armstrong, P. W., Westerhout, C. M., Van de Werf, F., Califf, R. M., Welsh, R. C., Wilcox, R. G., and Bakal, J. A. (2011). Refining clinical trial composite outcomes: An application to the assessment of the safety and efficacy of a new thrombolytic–3 (assent-3) trial. *American heart journal*, 161(5):848–854.

Cai, T., Tian, L., Uno, H., Solomon, S., and Wei, L. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika*, 97(2):389–404.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Ghosh, D. and Lin, D. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics*, 59(4):877–885.

Gilbert, P. B., Wei, L., Kosorok, M. R., and Clemens, J. D. (2002). Simultaneous inferences on the contrast of two hazard functions with censored observations. *Biometrics*, 58(4):773–780.

Glasziou, P., Simes, R., and Gelber, R. (1990). Quality adjusted survival analysis. *Statistics in medicine*, 9(11):1259–1276.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13.

Huang, C.-Y. and Wang, M.-C. (2011). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*.

Kalbfleisch, J. D. and Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika*, 68(1):105–112.

Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association*, 82(400):1169–1176.

Li, Q. and Lagakos, S. (1998). Use of the wei–lin–weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine*, 16(8):925–940.

Lin, D. (2003). Regression analysis of incomplete medical cost data. *Statistics in medicine*, 22(7):1181–1200.

Lin, D., Wei, L., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):pp. 1074–1078.

Liu, L., Wolfe, R. A., and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3):747–756.

Mao, L. and Lin, D. (2016). Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. *Biostatistics (Oxford, England)*, 17(2):390.

Park, Y. and Wei, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 90(3):717–723.

Parzen, M., Wei, L., and Ying, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics*, 24(3):309–314.

Rondeau, V., Gonzalez, J. R., Yassin Mazroui, A., Mauguen, A. D., Laurent, A., and Rondeau, M. V. (2013). Package frailtypack.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4):708–721.

Royston, P. and Parmar, M. K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine*, 30(19):2409–2421.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2):363–369.

Tian, L., Zhao, L., and Wei, L. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2):222–233.

Trinquart, L., Jacot, J., Conner, S. C., and Porcher, R. (2016). Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, page JCO642488.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22):2380–2385.

Uno, H., Wittes, J., Fu, H., Solomon, S. D., Claggett, B., Tian, L., Cai, T., Pfeffer, M. A., Evans, S. R., and Wei, L.-J. (2015). Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Annals of internal medicine*, 163(2):127–134.

Wang, M.-C. and Chiang, C.-T. (2002). Non-parametric methods for recurrent event data with informative and non-informative censorings. *Statistics in medicine*, 21(3):445–456.

Wang, M.-C. and Huang, C.-Y. (2014). Statistical inference methods for recurrent event processes with shape and size parameters. *Biometrika*, page asu016.

Wang, M.-C., Qin, J., and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):1057–1065.

Wei, L., Lin, D., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073.

Zhao, L., Tian, L., Uno, H., Solomon, S. D., Pfeffer, M. A., Schindler, J. S., and Wei, L. J. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*, 9(5):570–577.

Zucker, D. M. (1998). Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442):702–709.