Research Article

Statistics in Medicine

Received 3 July 2015, Accepted 6 March 2016 Published online 5 April 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6950

A rank test for bivariate time-to-event outcomes when one event is a surrogate

Pamela A. Shaw[a*†] and Michael P. Fay[b]

In many clinical settings, improving patient survival is of interest but a practical surrogate, such as time to disease progression, is instead used as a clinical trial's primary endpoint. A time-to-first endpoint (*e.g.*, death or disease progression) is commonly analyzed but may not be adequate to summarize patient outcomes if a subsequent event contains important additional information. We consider a surrogate outcome very generally as one correlated with the true endpoint of interest. Settings of interest include those where the surrogate indicates a beneficial outcome so that the usual time-to-first endpoint of death or surrogate event is nonsensical. We present a new two-sample test for bivariate, interval-censored time-to-event data, where one endpoint is a surrogate for the second, less frequently observed endpoint of true interest. This test examines whether patient groups have equal clinical severity. If the true endpoint rarely occurs, the proposed test acts like a weighted logrank test on the surrogate; if it occurs for most individuals, then our test acts like a weighted logrank test on the true endpoint. If the surrogate is a useful statistical surrogate, our test can have better power than tests based on the surrogate that naively handles the true endpoint. In settings where the surrogate is not valid (treatment affects the surrogate but not the true endpoint), our test incorporates the information regarding the lack of treatment effect from the observed true endpoints and hence is expected to have a dampened treatment effect compared with tests based on the surrogate alone. Published 2016. This article is a U.S. Government work and is in the public domain in the USA.

Keywords: bivariate survival; composite outcome; interval censoring; semi-competing risks; surrogate endpoints; survival analysis

1. Introduction

Challenges can arise in the analysis of clinical trials with surrogate outcomes when both a surrogate and true endpoint of interest may be observed, but the true endpoint is observed infrequently. Here, we use an informal definition of a surrogate endpoint that is often necessary in practice; namely, an endpoint that is practical to study in an early phase trial and one for which an efficacious treatment would be expected to also be efficacious for the true endpoint. We refer to this type of surrogate as a *working surrogate*, to distinguish it from that of a validated or true surrogate by the usual Prentice or other statistical criteria [1].

Our motivating setting is a randomized treatment trial for patients with multiple drug-resistant tuberculosis. Because of the length of treatment, generally 2 years or longer, early-phase clinical trials are typically done on a working surrogate marker, such as time to sputum conversion (first negative sputum culture). For patients with multiple drug-resistant tuberculosis, the risk of death even for a relatively short trial is not negligible. Treating death as censoring for time-to-conversion could lead to a misleading analysis of the treatment benefit. One could choose to ignore the surrogate event once the true endpoint is observed, but this could also lead to the loss of potentially useful information in a setting where there is a lot of censoring for both events. Rather than limiting the analysis in this setting to the timing of a single event, we propose a method that uses information from the bivariate survival distribution.

[a] *Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.*

[b] *Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, U.S.A.*

[*] *Correspondence to: Pamela Shaw, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.*

[†] *E-mail: shawp@upenn.edu*

When the true and surrogate endpoints are both detrimental events, like death and cancer relapse, it often makes sense to consider the time-to-first and redefine the endpoint as disease-free or progression-free survival. This traditional approach does not makes sense when the event of interest is a failure (death) and the surrogate is a success (sputum conversion), or vice versa. This setting is also different from the traditional competing risk problem where, typically, one considers the time to the first of a collection of events (such as death, stroke, heart failure, etc.) whose occurrence has the same connotation of success or failure [2]. In the cancer setting, a subject may be observed to go into remission (success) and then die from disease or treatment-related toxicities (failure). Adding to the complexity, different types of events may not be of equal interest or importance. Once a competing true endpoint like death is observed, the other event may no longer be of interest. Conversely, observation of the surrogate typically does not preclude observing the true event. The proposed method is able to address a composite of beneficial and detrimental endpoints and the potential asymmetry in the importance of the two events to the evaluation of a patient's outcome.

Much of the existing literature for analyzing multiple endpoints considers issues of testing multiplicity or how to efficiently combine univariate test statistics across multiple outcomes of equal importance and did not consider event-time outcomes [3–6]. DiRienzo and DeGruttola [7] compared strategies for a bivariate failure time endpoint, where both endpoints are surrogates and considered methods that estimated an average effect and an optimal omnibus test across the multiple endpoints. These methods treated the two surrogate endpoints as having equal importance and did not consider the scenario where one endpoint might be subject to heavy censoring. For strategies that combine univariate test statistics by inversely weighting according to the variance, a seldom-observed true endpoint would naturally be down-weighted relative to a more frequently observed surrogate [3, 7]. Finkelstein and Schoenfeld [8] present a conditional expected score test that estimates a joint treatment effect for two failure time events that are ordered in time, focusing in particular on disease progression and death, but their analysis also treats both endpoints as equally important and assumes the same treatment effect for both endpoints.

We propose a rank-based test for a group comparison of outcomes that considers the treatment efficacy on clinical severity and incorporates the bivariate event times of a short-term surrogate endpoint and long-term true endpoint of interest. The rank test permits a flexible way to assess patient outcome, allowing a setting-specific disease severity function. In the settings of interest, an observed true endpoint completely determines the clinical severity. The proposed method also addresses bivariate interval-censoring. When a subject is censored after an observed surrogate success and prior to the end of the trial or observation of the true outcome, this introduces a novel type of interval censoring for the proposed measure of disease severity. We present the asymptotic properties and conditions for which the proposed test would be asymptotically equivalent to a weighted logrank test on the event of interest. We further illustrate the method with a data example.

2. Preliminaries

2.1. Bivariate likelihood

Let $\mathbf{T}_i = (T_{1i}, T_{2i})$ denote paired event times that follow the bivariate distribution $F_{\mathbf{T}}(\mathbf{t}) = F_{\mathbf{T}}(t_1, t_2) = P(T_1 \leqslant t_1, T_2 \leqslant t_2)$, for $i = 1, \ldots, n$. In our setting, T_{1i} is the surrogate event time and T_{2i} is the true endpoint. Suppose $\mathbf{T}_i$ is censored according to some independent process and known only to be in the region $R_i = (l_{1i}, r_{1i}] \times (l_{2i}, r_{2i}]$. The likelihood for n independent observations is thus

$$L = \prod_{i=1}^{n} P_{\mathbf{T}}(R_i) = \prod_{i=1}^{n} \{F_{\mathbf{T}}(r_{1i}, r_{2i}) - F_{\mathbf{T}}(r_{1i}, l_{2i}) - F_{\mathbf{T}}(l_{1i}, r_{2i}) + F_{\mathbf{T}}(l_{1i}, l_{2i})\}.$$

We will use the bivariate non-parametric maximum likelihood estimator (NPMLE) for $F_{\mathbf{T}}$ in the methods that follow, as it does not rely on parametric assumptions; however, other estimators could be used. The NPMLE has been well studied [9–13] and shown to be strongly consistent for bivariate interval-censored data [14, 15]. Unlike the univariate case, the bivariate NPMLE can have mixture nonuniqueness [12], where the optimal distribution of mass across the regions in the support of the NPMLE may not be unique. Conditions that guarantee mixture uniqueness of the NPMLE are easily checked [12]. We assume that the NPMLE is mixture unique but return to this issue in the discussion.

Assume a discrete set of possible assessments at $\{a_{11}, ..., a_{1J}\}$ for T_1 and $\{a_{21}, ..., a_{2K}\}$ for T_2, with $0 \equiv a_{10} < a_{11} < \ldots < a_{1J} < \infty \equiv a_{1,J+1}$ and $0 \equiv a_{20} < a_{21} < \ldots < a_{2K} < \infty \equiv a_{2,K+1}$. We allow for mixed-case interval censoring, where each individual may be assessed on some subset of these assessment times. Asymptotically, this yields a finite number of regions for the finest assessment grid in which event times can be observed to reside and the NPMLE will be consistent on $\mathcal{A}$, the set of bivariate assessment times that have positive probability [14]. Let $u_j = (a_{1,j-1}, a_{1j}]$ for $j = 1, 2, \dots, J+1$, and $v_k = (a_{2,k-1}, a_k]$ for $k = 1, \dots, K+1$. For notational convenience only, we assume the same maximum assessment time τ for both events. Here, τ could be the intended length of follow-up for a clinical trial. Discrete assessment times impose no practical limitation because J and K can be as large as needed and are natural for the setting where the occurrence of an event is assessed periodically. Although it is customary to assume survival time is known precisely or right censored, it is frequently rounded to the nearest day and hence can also be considered discrete.

2.2. *Group-level tests for interval censored data*

Fay [16] presents a general form of the weighted logrank test for interval censored data of a univariate outcome. Our proposed test will be an extension of this test for a univariate severity score that incorporates information from the bivariate outcome. We review this test here and establish notation. Let $R_i = (l_i, r_i]$ be the observed event region for individual i, and let $S(t;\beta)$ be the survival distribution function for the event time T under the semi-parametric grouped continuous model. For a binary treatment indicator, z_i, the efficient score statistic is $U = \sum_{i=1}^{n} z_i \left\{\hat{S}'(l_i) - \hat{S}'(r_i)\right\} \left\{\hat{S}(l_i) - \hat{S}(r_i)\right\}^{-1} \equiv \sum_{i=1}^{n} z_i c_i$, where $\hat{S}'(t)$ is the derivative of $S(t;\beta)$ with respect to β evaluated at $\beta = 0$, and $\hat{S}$ is the NPMLE for S under the null. We write c_i as $r(R_i, \hat{F}_\mathbf{T})$, where $\hat{F}_\mathbf{T}$ is the NPMLE for F_T, to emphasize that it can be viewed as a ranking function. We assume the proportional odds model (i.e., logistic error), for which $c_i = r(R_i, \hat{F}_\mathbf{T}) = 1 - \hat{F}_\mathbf{T}(l_i) - \hat{F}_\mathbf{T}(r_i)$. The resulting permutation test is equivalent to the generalized Wilcoxon rank test of Peto and Peto [17]. The Wilcoxon test is natural for our setting, where ranking individuals according to clinical severity is of interest. See Fay [16] for further details and other choices for $r(R_i, \hat{F}_\mathbf{T})$. The k-group test can be similarly derived as a quadratic form using the score vector [16].

For the succeeding proposed method, it will be convenient to rewrite U. Define $u_j = (a_{j-1}, a_j]$ with $a_j \in \mathcal{A}$ for $j \in \{1, \dots, J\}$, the set of J assessment times, $a_0 \equiv 0$ and $a_{J+1} \equiv \infty$. Let α_{ij} be the indicator that $u_j \subset R_i$, for $i = 1, \dots, n$ and $j = 1, \dots, J+1$. Then

$$
\begin{aligned}
U &= \sum_{i=1}^{n} z_i \sum_{j=1}^{J+1} \frac{\alpha_{ij}\left\{\hat{S}'(a_{j-1}) - \hat{S}'(a_j)\right\}}{\sum_{h=1}^{J+1} \alpha_{ih}\left\{\hat{S}(a_{h-1}) - \hat{S}(a_h)\right\}} \\
&= \sum_{i=1}^{n} z_i \sum_{j=1}^{J+1} \left[\frac{\alpha_{ij}\left\{\hat{S}(a_{j-1}) - \hat{S}(a_j)\right\}}{\sum_{h=1}^{J+1} \alpha_{ih}\left\{\hat{S}(a_{h-1}) - \hat{S}(a_h)\right\}}\right] \left[\frac{\hat{S}'(a_{j-1}) - \hat{S}'(a_j)}{\hat{S}(a_{j-1}) - \hat{S}(a_j)}\right].
\end{aligned}
$$

From this expression, we see that U can also be written as

$$
\sum_{i=1}^{n} z_i \sum_{u_j \in R_i} \frac{\hat{P}_T(u_j) r(u_j, \hat{F}_T)}{\hat{P}_T(R_j)}, \tag{1}
$$

where $\hat{P}_T(u_j) = \hat{P}_T((a_{j-1}, a_j]) = \hat{F}_T(a_j) - \hat{F}_T(a_{j-1})$ estimates the probability that the event occurred in $(a_{j-1}, a_j]$ and $r(u_j, \hat{F}_T) = 1 - \hat{F}(a_{j-1}) - \hat{F}(a_j)$. The ith individual's contribution to the usual score statistic can be seen as a weighted average of the function $r(u_j, \hat{F}_T)$.

3. Proposed method

3.1. *Defining clinical severity*

The goal in many clinical research studies is to examine whether patients are better under treatment A or treatment B. When T_1 is a surrogate for T_2, it may make sense to have the severity completely determined by T_2 when it is observed and otherwise determined by T_1. We consider this special case in detail, as it is relevant for the setting that motivated this work and yields useful asymptotic properties. In this case, we will see that the proposed method reduces to a weighted summary statistic based on the two individual failure times. In other settings, T_1 may contain auxiliary information not provided by T_2 and a composite severity score that assigns weights to the different event times may be of interest. The severity score Q is used in a rank test, so only the ordering induced by Q is important. So long as the clinical severity score is quantified by a bounded real-valued function $Q(\mathbf{T})$, there will be sufficient regularity for the limiting behavior of the proposed test.

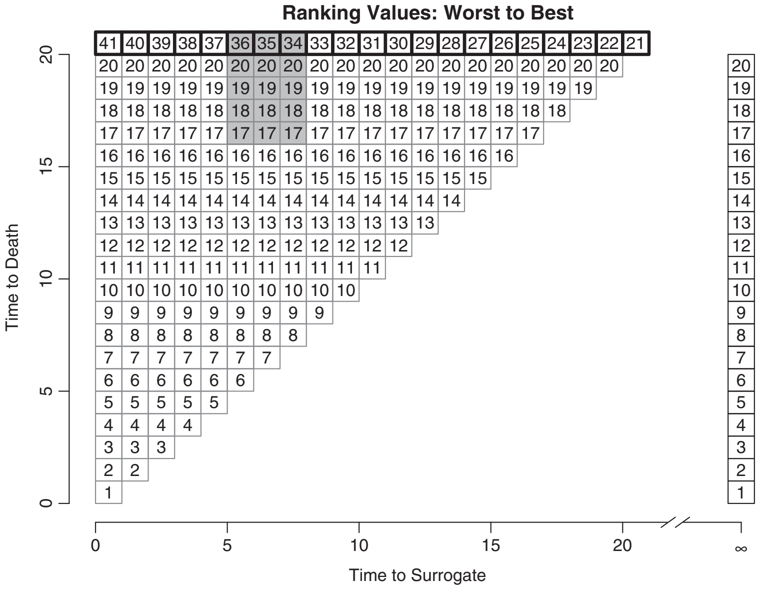
3.1.1. Surrogate good/true endpoint bad. Suppose a larger value of T_2 (survival) indicates less severe disease, but larger T_1 (surrogate) indicates more severe disease. Consider the following severity score Q, where high values of Q are associated with less severe disease.

Case 1 (No loss to follow-up)
In this case, everyone is observed until death or until the end of the trial, at time τ. For $\mathbf{t} = (t_1, t_2)$, define $Q(\mathbf{t}) \equiv Q(t_1, t_2)$ as

$$\begin{aligned} Q(t_1, t_2) &= t_2, \text{ if } t_2 \leqslant \tau \\ &= 2(\tau + 1) - t_1, \text{ if } t_2 > \tau. \end{aligned}$$

Thus, all survivors are scored as less severe than those who die. Later death times are scored as less severe; for survivors, earlier surrogate event times are less severe. When ranking severity, an event time that is censored at time τ can be considered as happening at a time ϵ beyond τ; we assume $\epsilon = 1$. The value of ϵ is arbitrary and serves only to break ties in the severity for the cases where a subject is observed to die at time τ, survive without converting by time τ, or convert on day τ. In these cases Q takes on the values τ, $\tau + 1$, and $\tau + 2$, respectively. Figure 1 illustrates the assessment grid, assuming the possible inspection times are the positive integers up to $\tau = 20$ for both the surrogate and survival times and displaying the severity score $Q(a_{1j}, a_{2k})$ in the square defined by $u_j \times v_k$. Thus, events can happen at time 1, ..., 20, and the possible values for Q range from 1 (subject died at week 1) to 41 (subject converted at week 1 and survived the length of the trial). For those observed to die before the surrogate success, the possible severity ranks are shown at $T_1 = \infty$. Someone who survives the trial without converting has a severity score of $Q(21, 21) = 21$. The next best score is for someone who converted on the 20th week, Q(21,20) = 22, etc.

Case 2 (Loss to follow-up)
Now suppose there is loss to follow-up for the ith subject at time L, where $0 < L < \tau$. $Q(\mathbf{T})$ is interval-censored in $(L, 2\tau - L + 2)$ if $\min(T_1, T_2)$ is known to be greater than L and interval censored in $(L, \tau] \cup [2\tau - a_{1j} + 2, 2\tau - a_{1,j-1} + 2)$ if T_1 is known to be in $(a_{1,j-1}, a_{1,j}]$, where $a_{1,j} < L$. Figure 1 shows the interval-censored region for $Q(\mathbf{T})$ for an individual with T_1 interval-censored $(5, 8]$ and T_2 right censored at time 16.



41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	
20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20		20
19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19			19
18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18				18
17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17					17
16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16						16
15	15	15	15	15	15	15	15	15	15	15	15	15	15	15							15
14	14	14	14	14	14	14	14	14	14	14	14	14	14								14
13	13	13	13	13	13	13	13	13	13	13	13	13									13
12	12	12	12	12	12	12	12	12	12	12	12										12
11	11	11	11	11	11	11	11	11	11	11											11
10	10	10	10	10	10	10	10	10	10												10
9	9	9	9	9	9	9	9	9													9
8	8	8	8	8	8	8	8														8
7	7	7	7	7	7	7															7
6	6	6	6	6	6																6
5	5	5	5	5																	5
4	4	4	4																		4
3	3	3																			3
2	2																				2
1																					1

Figure 1. The assessment grid defined by positive-integer inspection times up to $\tau = 20$. The severity score $Q(a_{1j}, a_{2k})$ is shown in the square $(a_{1,j-1}, a_{1,j}] \times (a_{2,k-1}, a_{2,k}]$. The shaded region indicates (T_1, T_2) occurs in $(5, 8] \times (16, \infty)$.

3.1.2. Surrogate bad/true endpoint worse. It is straightforward to modify Q for other settings, where earlier times are clinically either both worse or both better for the two endpoints. Suppose earlier times are worse.

Case 1 (No loss to follow-up)
We define

$$\begin{aligned} Q(t_1, t_2) &= t_2, \text{ if } t_2 \leqslant \tau \\ &= \tau + t_1, \text{ if } t_2 > \tau, \end{aligned}$$

where the severity for someone who survives τ with no surrogate event is defined by $2\tau + 1$. Once again, all the survivors (of the true endpoint) will be ranked better than those who died (have an observed true endpoint).

Case 2 (Loss to follow-up)
Now suppose there is loss to follow-up for the ith subject at time L, where $0 < L < \tau$. $Q(\mathbf{T})$ is interval censored in $(L, \tau] \cup (\tau + L, 2\tau + 1]$ if $\min(T_1, T_2)$ is known to be greater than L and interval censored in $(L, \tau] \cup (\tau + a_{1j-1}, \tau + a_{1,j}]$ if T_1 is known to be in $(a_{1,j-1}, a_{1,j}]$, where $a_{1,j} < L$.

3.2. *Estimating the severity distribution function*

In Section 3.1.1, a univariate severity score $Q(\mathbf{T})$ is defined as a function of the interval-censored bivariate outcome $\mathbf{T}$. Given the bivariate distribution function $F_{\mathbf{T}}$, a cumulative distribution function F_Q for the severity score is induced. We propose the use of the bivariate NPMLE $\hat{F}_{\mathbf{T}}$. Let $M = \{M_\ell\}$ be the set of all rectangular regions (of maximal intersection) M_ℓ in the support of $\hat{F}_{\mathbf{T}}$. Recall from Section 2.1, the discrete set of possible assessment times allow us to think of each M_ℓ as the union of one or more of the non-overlapping, irreducible regions $u_j \times v_k$ from the underlying assessment grid. For the proposed method, we will need to specify the distribution of mass across these irreducible regions in each M_ℓ. We propose the probability mass within M_ℓ be distributed uniformly. Other possible estimators could include a semi-parametric approach that considers kernel type estimates of this distribution based on the neighboring density distribution. In Section 4, we establish that asymptotically, under certain regularity conditions, the distribution of mass across the grid units within $M_\ell \in M$ is inconsequential. We define $\hat{F}_Q$ as

$$\hat{F}_Q(q) = \hat{P}(Q \leqslant q) = \sum_{M_\ell \in M} \sum_{(a_{1j}, a_{2k}) \in M_\ell \cap \mathcal{A}} I\{Q(a_{1j}, a_{2k}) \leqslant q\} \frac{\hat{P}_{\mathbf{T}}(M_\ell)}{\#\{M_\ell \cap \mathcal{A}\}}, \tag{2}$$

where $\#\{M_\ell \cap \mathcal{A}\}$ is the number of grid units $u_j \times v_k$ in M_ℓ. Note, by definition, each maximal clique contains at least one point on the assessment grid, thus $\#\{M_\ell \cap \mathcal{A}\} > 0$. From the definition of Q in Section 3.1.1, there are $J + K + 1$ possible severity scores $0 < q_1 < .... < q_{J+K+1} < \infty$ in the range of Q.

3.3. *Proposed hypothesis test*

We propose a group-level test, based on the ranks of clinical severity Q, that uses the NPMLE for bivariate survival distribution $F_{\mathbf{T}}$ and accommodates the interval-censoring structure of the data. We show that this test reduces to a weighted combination of two test statistics for the special case of surrogacy considered in Section 3.1.

The usual two-sample test for univariate interval-censored data, described in Section 2.2, is not well-suited for an analysis of the clinical severity score Q for two reasons. Firstly, it assumes a contiguous interval for the interval-censored event region, and secondly, it does not fully take advantage of the bivariate information in the observed data $\{\mathbf{T}_i\}$. The possible severity scores for an individual with a right-censored true endpoint and a surrogate known to occur will be noncontiguous, as demonstrated in Figure 1. There could also be multiple paired event times with the same severity Q but different weights in the bivariate distribution $\hat{F}_{\mathbf{T}}$. To account for this, we propose the ith individual's contribution to the test statistic be a weighted sum of the ranking function of the possible severity scores that Q assigns to the set of irreducible assessment regions $u_j \times v_k$ in the individual's R_i (here a bivariate event region) and whose weights are based on the conditional distribution given $\mathbf{T} \in R_i$. Define $Q^*(u_j \times v_k) = Q(a_{1j}, a_{2k})$. Suppose $Q(a_{1j}, a_{2k}) = q_l$, that is, the lth largest possible severity score, then one has the ranking function $r\{Q^*(u_j \times v_k), \hat{F}_Q\} = 1 - \hat{F}_Q(q_l) - \hat{F}_Q(q_{l-1})$. Thus, the proposed statistic is

$$U = \sum_{i=1}^{n} z_i c_i = \sum_{i=1}^{n} z_i \sum_{u_j \times v_k \subset R_i} \frac{\hat{P}_T(u_j \times v_k) r\{Q^*(u_j \times v_k), \hat{F}_Q\}}{\hat{P}_T(R_i)}. \tag{3}$$

A p-value can be obtained by performing a permutation test.

To understand the properties of this test statistic, it will be convenient to rewrite Equation (3) as a weighted sum of two test statistics. Let α_{ijk} be the indicator that $u_j \times v_k$ is in the ith event region R_i, and let $\hat{p}_{jk}$ be the mass assigned to $u_j \times v_k$ by $\hat{F}_T$. Define $\hat{w}_i = \sum_{j=1}^{J+1} \sum_{k=1}^{K} \frac{\alpha_{ijk}\hat{p}_{jk}}{\sum_{l=1}^{J+1}\sum_{m=1}^{K+1} \alpha_{ilm}\hat{p}_{lm}}$, then $\hat{w}_i = \hat{P}_{\mathbf{T}}(T_{2i} \leqslant \tau | \mathbf{T}_i \in R_i)$ and the ith individual's contribution to the score statistic in (3) can be seen as $c_i z_i$, where

$$\begin{aligned} c_i &= \sum_{j=1}^{J+1}\sum_{k=1}^{K} \frac{\alpha_{ijk}\hat{p}_{jk} r\left\{Q^*(u_j \times v_k), \hat{F}_Q\right\}}{\sum_{l=1}^{J+1}\sum_{m=1}^{K+1} \alpha_{ilm}\hat{p}_{lm}} + \sum_{j=1}^{J+1} \frac{\alpha_{ij(K+1)}\hat{p}_{j(K+1)} r\left\{Q^*(u_j \times v_{K+1}), \hat{F}_Q\right\}}{\sum_{l=1}^{J+1}\sum_{m=1}^{K+1} \alpha_{ilm}\hat{p}_{lm}} \\ &= \hat{w}_i \sum_{j=1}^{J+1}\sum_{k=1}^{K} \frac{\alpha_{ijk}\hat{p}_{jk} r\left\{Q^*(u_j \times v_k), \hat{F}_Q\right\}}{\sum_{l=1}^{J+1}\sum_{m=1}^{K} \alpha_{ilm}\hat{p}_{lm}} + (1 - \hat{w}_i) \sum_{j=1}^{J+1} \frac{\alpha_{ij(K+1)}\hat{p}_{j(K+1)} r\left\{Q^*(u_j \times v_{K+1}), \hat{F}_Q\right\}}{\sum_{l=1}^{J+1} \alpha_{il(K+1)}\hat{p}_{l(K+1)}}. \end{aligned}$$

Then one has

$$\begin{aligned} n^{-1}U = &\left(\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i\right) \left[\sum_{i=1}^{n} \frac{\hat{w}_i}{\sum_{v=1}^{n} \hat{w}_v} \frac{\sum_{k=1}^{K}\sum_{j=1}^{J+1} \alpha_{ijk}p_{jk} r\{Q^*(u_j \times v_k), \hat{F}_Q\}}{\sum_{m=1}^{K}\sum_{l=1}^{J+1} \alpha_{ilm}p_{lm}} z_i\right] \\ &+ \left\{\frac{1}{n}\sum_{i=1}^{n}(1 - \hat{w}_i)\right\} \sum_{i=1}^{n} \frac{(1 - \hat{w}_i)}{\sum_{v=1}^{n}(1 - \hat{w}_v)} \frac{\sum_{j=1}^{J+1} \alpha_{ij(K+1)}p_{j(K+1)} r\{Q^*(u_j \times v_{K+1}), \hat{F}_Q\}}{\sum_{l=1}^{J+1} \alpha_{il(K+1)}p_{l(K+1)}} z_i \\ &\equiv \hat{P}(T_2 \leqslant \tau) \cdot U_2 + \{1 - \hat{P}(T_2 \leqslant \tau)\} \cdot U_1. \end{aligned} \tag{4}$$

U is thus a weighted sum of two statistics: U_2, which can be viewed as a re-weighted version of the univariate score statistic (1) for interval-censored T_2, with weights equal to an individual's probability of having a death before τ relative to the rest of the cohort; and U_1, which can similarly be seen as a re-weighted statistic based on a linearly transformed surrogate for survivors ($2\tau - T_1 + 2$ for Q defined as in 3.1.1 and $\tau + T_1$ for Q defined as in 3.1.2), with weights equal to person i's relative probability of having $T_2 > \tau$. The estimated $P(T_2 \leqslant \tau)$ determines the weight given to U_2. Note, if T_2 was censored for all subjects, the necessary quantities in (4) would still be estimable. In this case, $\hat{P}(T_2 \leqslant \tau) = 0$. The maximal cliques would be univariate intervals and (4) reduces to the univariate score statistic based on the NPMLE for the univariate marginal of T_1. Similarly, if death was always observed, then (4) reduces to a univariate statistic based on the NPMLE for the univariate marginal of T_2.

4. Properties

We describe the regularity conditions that allow for a valid permutation test. We also provide an intuitive argument for why the proposed test is approximately equivalent to the usual score test for interval-censored survival data for the true endpoint alone (T_2), when the trial is allowed to run long enough that T_2 will have occurred on everyone. Note, this is not the same as assuming there is no censoring. The result holds under interval or right-censoring due to missed assessments; we assume the trial is long enough that there will be no survivors at the end of the trial, that is, $P[T_2 > \tau] = 0$. Formal proofs of both results are provided in the Supporting Information.

First, we require the following standard assumption for censored time-to-event data, as stated in $A1$.

($A1$) The inspection process is jointly independent of the treatment Z and the bivariate failure time $\mathbf{T} = (T_1, T_2)$, and $(\mathbf{T}_i, Z_i)$ are independent and identically distributed (i.i.d.) for $i = 1, \ldots, n$.

With this assumption, we have the following result, which establishes the validity of the permutation test for the proposed test statistic in Equation (3).

Theorem 1
Given assumption A1, the severity function in Section 3.1, and U defined in Equation (3), one has the two-sample permutation test using $W = U$ as a test statistic is valid, that is, maintains the type I error rate under the null hypothesis of no treatment effect.

For our second result, we need two additional assumptions, formally stated in the succeeding text. Assumption $A2$ assumes the trial is long enough for all individuals to have the true event. Assumption $A3$ sets up the semi-competing risk problem, which is reasonable when the event of interest is death, since even if a surrogate event is indicative of a cure, individuals may die of other causes. The results can also be readily adapted for non-semicompeting risk data.

($A2$) Let $a_{2K} = \tau$ be such that $\tau < \infty$ and $P(T_2 > a_{2K}) = 0$.

($A3$) Let A_i be the set of possible bivariate inspection times for the ith individual. Denote the possible inspection times as $\{a_{1j}\}_{j=1}^{J}$ for T_1 and $\{a_{2k}\}_{k=1}^{K}$ for T_2. Assume the A_i are i.i.d. such that each pair (a_{1j}, a_{2k}) consistent with $T_1 < T_2$ has positive probability of occurring, and only these pairs. Call this set $\mathcal{A}$.

We now have the following result, which establishes that our proposed test statistic reduces to a univariate test on T_2.

Theorem 2
Given assumptions $A2$–$A3$, severity function defined as in Section 3.1.1 or 3.1.2, U defined in equation (3), and U_2 in Equation (4), one has $n^{-1}U - U_2 \xrightarrow{a.s.} 0$.

The full proof is given in the Supporting Information. Briefly, with the assumed regularity, the bivariate NPMLE will be strongly consistent at any observation time $\mathbf{a} \in \mathcal{A}$ [14]. The test statistic (3) relies on the estimated distribution functions only through $\mathbf{a} \in \mathcal{A}$. Furthermore, because all individuals will either be right-censored or have an observed event T_2 at a time less than a_{2K}, the NPMLE will assign no weight to regions for which $T_2 > a_{2K}$. Thus, $\hat{P}(T_2 \leqslant \tau) = 1$, and by Equation (4), one has $n^{-1}U = U_2$, which gives the result. One can also show that $n^{-1}U$ is approximately equivalent to the usual univariate weighted logrank statistic for T_2, with the only difference being that the bivariate NPMLE is used to estimate the marginal for T_2. It can be shown similarly that for rarely occurring T_2, the proposed test will be approximately equal to the univariate test for interval censored T_1. For this result, we rely on the approximate equivalence of F_{T_1} and $F_{T_1|T_2>\tau}$ for $P(T_2 > \tau) \approx 1$ and on the invariance of the chosen generalized Wilcoxon ranking function $r(\cdot, F)$ under monotonic transformations.

5. Simulation studies

We consider a hypothetical phase II clinical trial examining a novel regimen with potential toxicity concerns. The Linezolid Trial for XDR-TB is one such trial, which studied the efficacy of adding linezolid immediately versus after an 8-week delayed start to the background regimen on the working surrogate marker of time to sputum culture conversion from positive to negative [18]. We numerically simulated a trial similar to this setting, where T_1 is time to sputum conversion and T_2 is survival, censored at 20 weeks. We also simulate a long-term trial, where both endpoints are subject to less censoring. We assume a treatment benefit for the surrogate, that is, quicker times to sputum conversion, but vary the effect on survival.

We examine the performance of several rank-based methods to evaluate treatment efficacy, including (i) the *naive* method: a generalized Wilcoxon test [17] for the difference in time to sputum conversion between treatment groups that treats death as a censoring event; (ii) the *weak* method: a generalized Wilcoxon test for the survival endpoint; (iii) the *death* penalty: which assigns anyone who dies the worst outcome and otherwise is the generalized Wilcoxon test for surrogate event; and (iv) the *proposed*, the rank-based test for the clinical severity score. The severity score of Section 3.1.1 is used. Analysis is done in R version 2.14 [19]. The Icens and MLEcens packages are used to find the regions of maximal intersection and the bivariate NPMLE, respectively. The ranking functions $r\left(Q^{*}(u_j \times v_k), F_Q\right)$ are calculated using the interval package, and the permutation tests were done with perm package [20]. For the permutation tests, we set the perm method="exact.mc", which then estimated the exact p-value using Monte Carlo simulation and a specified 10,000 iterations.

We generate the bivariate endpoint $\mathbf{T} = (T_1, T_2)$ using hierarchical relative models, where the surrogate event time T_1 influences the survival time T_2 and both outcomes (T_1, T_2) can be influenced by treatment Z. We assume $\lambda_{T_1}(t_1) = \lambda_{01}(t_1)\exp(\alpha z)$ and $\lambda_{T_2}(t_2) = \lambda_{02}(t_2)\exp(\beta_1 t_1 + \beta_2 z)$. Dependence between T_1 and T_2 is thus modeled by the parameter β_1. T_1 can be thought of as a potentially latent variable associated with the underlying disease process, which may not be observed. We do not assume T_1 is a true surrogate, in the sense of Prentice [1], and allow the treatment to have an independent effect on death not captured by T_1 ($\beta_2 \neq 0$). A treatment survival benefit mediated through the surrogate is modeled with a positive α and β_1. A negative (positive) β_2 allows for an independent beneficial (detrimental) effect of treatment on survival. For simplicity, T_1 and T_2 are assumed to be reported on the same, weekly, time scale. Baseline survival distributions for T_1 and T_2 endpoints are simulated as exponential $(\gamma_j), j = 1, 2$ and then rounded to integer time.

In scenario 1, the relative and baseline hazard parameters were chosen so that the naive Wilcoxon test for the surrogate has roughly 80% power for 50 individuals per arm in the short trial, and this test for efficacy based on the true and surrogate endpoint had similar power (roughly 85%) in the long trial, namely, $(\alpha, \beta_1, \beta_2) = (0.8, 1.5, -0.5)$, $(\gamma_1, \gamma_2) = (20, 100)$, $\tau_{short} = 20$ and $\tau_{long} = 170$. Individuals were also subject to right censoring due to loss to follow-up, modeled with a 30% probability of dropout before τ and dropout time distributed uniformly over $(0, \tau)$. These parameters are associated with approximately 30% (70%) censoring for the surrogate (survival) event for the short trial and 20% censoring for both event times in the long trial. Results from the long trial for the *weak* (survival) endpoint can be considered the gold standard benchmark, in that they are results of a test of efficacy for the (impractical) trial based on the true endpoint of interest.

Table I shows the probability of rejecting for benefit for scenario 1. In the short trial, the naive and proposed methods yield very similar results, with the proposed having slightly more power as it was able to use information on the second endpoint. The *weak* method based on survival alone had low power (43.5%), due to the high level of censoring. The *death penalty* method, which treated all deaths with the same (worst) event time, did well in the short trial but performs relatively poorly in the long trial with only 70% power, due to the number of ties in the failure times induced by giving all deaths the same failure time. As expected, for the long trial, the proposed method provides similar results to the usual analysis of survival. Because the net treatment benefit for the survival endpoint and the surrogate were similar, the power for both these endpoints in the long trial is similar.

Table II shows the power for three scenarios with the same treatment effect for the surrogate ($\alpha = 0.8$), but vary the treatment benefit for survival to be (1) a weaker benefit than for the surrogate with a benefit only mediated through the surrogate, and roughly 20% power for the benchmark long-term trial results for survival ($\alpha = 0.8, \beta_1 = 1.5, \beta_2 = 0$); (2) useless, with a mortality hazard ratio of 1 for both indirect

Table I. Probability of rejecting for benefit by trial length, when there is efficacy for both surrogate and survival endpoints.

Estimators	Short	Long
Naive	79.8	83.8
Weak (survival)	43.5	86.2
Death penalty	84.6	70.0
Proposed	82.4	86.7

Table II. Probability of rejecting for benefit by trial length, when there is efficacy for a surrogate and different assumptions for treatment efficacy on survival.

	Weaker benefit		No effect		Harm	
Estimators	Short	Long	Short	Long	Short	Long
Naive	79.4	83.8	86.0	89.7	76.5	80.6
Weak (survival)	12.8	20.5	3.1	2.7	0.0	0.0
Death penalty	67.4	12.2	69.4	6.7	5.7	0.0
Proposed	56.9	21.2	52.1	3.6	0.5	0.0

and direct effects of treatment ($\alpha = 0.8, \beta_1 = 0, \beta_2 = 0$); and (3) harmful, by keeping the indirect benefit of treatment conveyed through shorter conversion terms and adding an independent, detrimental effect of treatment that outweighs the benefit mediated by the surrogate ($\alpha = 0.8, \beta_1 = 1.5, \beta_2 = 1$).

Consider first the weaker benefit scenario. In the short trial, the naive method has similar power as before, which is now more optimistic than what would be observed in the longer trial for survival. The proposed method provides some skepticism in the short term with a power of 56.9% and, for the long-term trial, has similar, only slightly elevated power of 21.2% compared with the 20.5% for the true survival endpoint. For the no treatment benefit scenario, the proposed method has less power than the naive analysis in the short-term and, for the long-term trial, has power only slightly above the nominal 2.5% one-sided type I error rate for survival. For the scenario with an overall harmful effect of treatment, the proposed method has almost no chance of rejecting for benefit in either the short or long trial. The naive analysis has a high probability of rejecting for benefit in all three scenarios, because the effect on the surrogate remains the same and death is treated as censoring. The death penalty method had similar results as the proposed for both the long and short trials of the no effect and harm scenarios. For this method, the loss of information is not an issue for the useless treatment scenario and is actually helpful in the harmful treatment scenario. The proposed method was the only method that provided reasonable answers in all scenarios.

We conducted a similar set of scenarios assuming an alternate model for the bivariate event times, namely, a linear transformation model where the log-transformed **T** conditioned on Z has a normal error distribution. Details of these simulations and results are presented in the Supporting Information. The relative performance of the methods was similar to those results presented here. In the Supporting Information, we also confirm the validity of the proposed test by performing simulations under the model above and the alternate survival model, assuming a null treatment effect and varying sample sizes. As expected, the proposed method, which relies on the exact permutation test, preserved the Type I error.

6. Studies of Left Ventricular Dysfunction trial

The Studies of Left Ventricular Dysfunction trials were randomized placebo-controlled trials of the efficacy of enalapril for the treatment and prevention of congestive heart failure to improve survival in patients with a weak left ventricular ejection fraction [21]. A key secondary aim was to assess the effect of treatment on hospitalization for congestive heart failure. In the treatment trial, enalapril was associated with a 16% decrease in mortality risk ($p = 0.0036$) and a 26% reduction in hospitalization for congestive heart failure ($p < 0.0001$). As an illustrative example for the proposed method, we present an analysis of these endpoints in a subset of 662 subjects with diabetes. Data and R code for the analysis are provided in the Supporting Information.

In order to summarize treatment efficacy on the patient's clinical severity, we consider both hospitalization (T_1) and mortality (T_2). For both endpoints, the sooner the event, the worse off the patient, so the time-to-first endpoint is appropriate here. We also consider a surrogate severity score similar to that of Section 3.1.2, so that earlier surrogate events (hospitalization) are ranked worse than later ones, while maintaining that survivors are ranked better than those who die. In this trial, not all subjects had the same length of follow-up. The subject with the longest follow-up determines τ. For survivors with a shorter follow-up time, say $u < \tau$, their severity score is then interval-censored between the value of Q that assumes they died at time $u + 1$ and the score that assumes they survived τ.

We compare the proposed test that examines the treatment effect on Q with the usual univariate logrank test for the treatment effect on: time to death, time to hospitalization (treating death as a censoring event), and the usual time-to-first of hospitalization or death. We also provide the p-values for the generalized Wilcoxon test. Table III presents the mortality and hospitalization outcomes by arm, as well as the usual Cox proportional hazards ratio and logrank test results for three time-to-event endpoints. For diabetic subjects, there was no survival benefit; however, the treatment had a 40% reduction in risk of hospitalization ($p < 0.0001$) and a 29% reduction in risk for the time to first endpoint, hospitalization or death, ($p = 0.0007$). One aspect of the mortality events that was not captured by the time-to-first endpoint is that hospitalizations on the enalapril arm were more likely to be followed by a death than the placebo arm. For the enalapril arm, 57/94 (61%) of hospitalizations were followed by a death, whereas on placebo 64/148 (43%) of the hospitalizations were followed by a death. The proposed rank test on the severity score has a p-value= 0.07 for the treatment group difference. Such a result helps quantify the level of evidence for an overall beneficial effect of enalapril, in a way that considers the timing of both events when they both occur rather than ignoring information on the more serious, second event.

Table III. Hospitalization, mortality, and time-to-first hospitalization or death (TTF) endpoints for subjects in SOLVD treatment trial with diabetes ($N = 662^*$).

	Enalapril ($N = 319$)		Placebo ($N = 343$)		Cox PH	Logrank	Generalized Wilcoxon
Endpoint	Yes	No	Yes	No	HR	*P*-value	*P*-value
Death	137	182	145	198	0.99	0.91	0.76
Hospitalization	94	225	148	195	0.60	< 0.0001	< 0.0001
TTF	174	145	229	114	0.71	0.0007	0.0007

[*]One of 663 diabetic subjects was excluded due to missing hospitalization event time.
TTF, time-to-first hospitalization or death.

7. Discussion

We considered a new two-sample test for bivariate time-to-event outcomes that relied on a ranking according to a clinical severity function. The proposed test has an advantage over the traditional time-to-first approach in that it can incorporate information on the complete clinical history of the patient. Additionally, if the time-to-first endpoint is a composite across events of disparate severity, ignoring this disparity can be misleading, particularly if there is an inconsistent treatment effect across these endpoints [22, 23].

In the examples we considered, the true endpoint of interest (survival) determined the clinical severity when it was observed. We saw that the proposed method had the nice property of increasing power to detect treatment benefit when there was an independent effect of treatment not captured by the surrogate and dampening this power when the treatment was less efficacious, or even harmful, for the survival endpoint. Additionally, the developed method allowed us to summarize patient severity considering both a positive surrogate (i.e., sputum conversion) and a detrimental true endpoint (i.e., death), a setting for which the traditional time-to-first endpoint did not make sense. The proposed test could be readily extended to consider other possible functions for the severity, as appropriate for the clinical setting. In some settings, the relative severity of different events may not be as straightforward. One could similarly perform a sensitivity analysis of different severity scores to see how the different test results agreed with expert evaluation of overall benefit. Although simplistic, such composite endpoints can be valuable aids in a complex decision process. Freedman *et al.* [24] suggest that the more complex the decision process becomes, the more decision bodies seem to prefer an objective rule to balance the weight of evidence.

The proposed score test relies on an NPMLE estimator for the bivariate survival distribution. This NPMLE is actually an equivalence class of functions defined by the probabilities assigned to the maximal intersection of the observed event regions. One analytic concern is that in some cases, there will not be a unique solution to the optimization problem that determines the probability assignments that maximize the likelihood. This type of distributional non-identifiability has been noted previously for the bivariate NPMLE. In our data example and numerical simulations, we found no examples of this non-uniqueness. Should this kind of non-identifiability arise in a real data example, one could do a sensitivity analysis or randomly choose among the NPMLEs, which would provide a valid test. An alternate solution would be to consider a parametric estimator for the bivariate survival function. Our proposed score test was presented for the two-sample case; however, it can be easily extended for the k-sample case by similarly adapting the usual log rank test for this setting.

In many disease settings, survival or cure are often of interest but are too impractical as primary endpoints. Surrogates are commonly selected as primary endpoints to enable trials to be of reasonable size and duration, particularly for early-phase trials. We proposed a novel test statistic for treatment efficacy that incorporates the relative severity of a surrogate and true endpoint of interest. The severity function approach is adaptable to many settings, as it can summarize not only disease severity across multiple event times but could also be used to assess treatment utility or account for quality of life. This method provides a useful alternative to the usual time-to-first event in settings where both event types are expected to occur during the trial. It may also be useful as a supportive composite outcome in a trial for which the primary outcome remains a traditional time to first endpoint.

References

1. Prentice R. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8**(4): 431–440.

2. Kabfleisch J, Prentice R. *The Analysis of Time-Failure Data* (2nd edn.) John Wiley and Sons, Inc.: New York, 2002.
3. Liu A, Li Q, Liu C, Yu K, Yu KF. A rank based test for comparison of multidimensional outcomes. *Journal of the American Statistical Association* 2010; **105**(490):578–587.
4. Huang P, Tilley B, Woolson R, Lipsitz S. Adjusting O'Brien's test to control type i error for the generalized nonparametric Behrens–Fisher problem. *Biometrics* 2005; **61**(2):532–539.
5. Pocock S, Geller N, Tsiatis A. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**:487–498.
6. O'Brien P. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**:1079–1087.
7. DiRienzo A, DeGruttola V. Design and analysis of clinical trials with a bivariate failure time endpoint, with application to AIDS Clinical Trials Group Study A5142. *Controlled Clinical Trials* 2003; **24**(2):122–134.
8. Finkelstein D, Schoenfeld DA. A joint test for progression and survival with interval-censored data from a cancer trial. *Statistics in Medicine* 2014; **33**:1981–1989.
9. van der Laan MJ. Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics* 1996; **24**(2):596–627.
10. Betensky R, Finkelstein D. A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine* 1999; **18**(22):3089–3100.
11. Prentice R. On non-parametric maximum likelihood estimation of the bivariate survivor function. *Statistics in Medicine* 1999; **18**(17-18):2517–2527.
12. Gentleman R, Vandal AC. Nonparametric estimation of the bivariate CDF for arbitrarily censored data. *Canadian Journal of Statistics* 2002; **30**(4):557–571.
13. Maathuis M. Reduction algorithm for the NPMLE for the distribution function of bivariate interval-censored data. *Journal of Computational and Graphical Statistics* 2005; **14**(2):352–362.
14. Yu S, Yu Q, Wong G. Consistency of the generalized MLE of a joint distribution function with multivariate interval-censored data. *Journal of Multivariate Analysis* 2006; **97**(3):720–732.
15. van der Vaart A, Wellner J. Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High Dimensional Probability II*, Giné E, Mason D, Wellner J (eds). Birkhäuser: Boston, 2000; 115–134.
16. Fay M. Comparing several score tests for interval censored data. *Statistics in Medicine* 1999; **18**(3):273–285.
17. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A* 1972; **135**:185–207.
18. Lee M, Lee J, Carroll MW, Choi H, Min S, Song T, Via LE, Goldfeder LC, Kang E, Jin B, Park H, Kwak H, Kim H, Heon HS, Jeong I, Joh JS, Chen RY, Olivier KN, Shaw PA, Follman D, Song SD, Lee JK, Lee DH, Kim CT, Dartois V, Park SK, Cho SN, Barry CE. Linezolid for treatment of chronic extensively drug-resistant tuberculosis. *New England Journal of Medicine* 2012; **367**(16):1508–1518.
19. R Development Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, 2011. Available from: http://www.R-project.org/ [Accessed on 25 March 2016], ISBN 3-900051-07-0.
20. Fay MP, Shaw PA. Exact and asymptotic weighted logrank tests for interval censored data: the interval R package. *Journal of Statistical Software* 2010; **36**(2):1–34. Available from: http://www.jstatsoft.org/v36/i02/ [Accessed on 25 March 2016].
21. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine* 1991; **325**:293–302.
22. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *British Medical Journal* 2010; **341**:1–7.
23. Kleist P. Composite endpoints for clinical trials: current perspectives. *International Journal of Pharmaceutical Medicine* 2007; **21**(3):187–198.
24. Freedman L, Anderson G, Kipnis V, Prentice R, Wang C, Rossouw J, Wittes J, DeMets D. Approaches to monitoring the results of long-term disease prevention trials: examples from the Women's Health Initiative. *Controlled Clinical Trials* 1996; **17**(6):509–525.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.