Published in final edited form as:

Stat Med. 2018 November 30; 37(27): 3918–3930. doi:10.1002/sim.7849.

Properties of Composite Time-to-First-Event versus Joint Marginal Analyses of Multiple Outcomes

Ionut Bebu^{a,*} and John M. Lachin^a

^aThe Biostatistics Center, Department of Epidemiology and Biostatistics, The George Washington University

Abstract

Many clinical studies (e.g., cardiovascular outcome trials) investigate the effect of an intervention on multiple event-time outcomes. The most common method of analysis is a so-called "composite" analysis of a composite outcome defined as the time to the first component event. Other approaches have been proposed, including the win ratio (or win difference) for ordered outcomes and the application of the Wei-Lachin test. Herein we assess the influence of the marginal and joint distributions of the component events, and their correlation structures, on the operating characteristics of these methods for the analysis of multiple events. The operating characteristics are investigated using a bivariate exponential model with a shared frailty, under which these properties are obtained in closed form.

While the composite time-to-first-event analysis provides an unbiased test of the hypothesis of equality of the distribution of the time to first event, we show that it can provide a biased test of the joint null hypothesis of equal marginal hazards when the correlation of event times differs between groups. The same applies to the Win Ratio. However, the operating characteristics of the Wei-Lachin or other tests of the joint equality of the marginal hazards are unaffected. Further, when the correlations are equal, the Wei-Lachin test is more powerful to detect a difference in marginal hazards than the composite analysis test. Careful consideration of the properties of the various methods for analysis of composite outcome measures are in order before adopting one as primary analysis in a clinical study.

Keywords

Composite outcome; event-time data; win ratio/difference

1. Introduction

In many settings, multiple time-to-event outcomes are combined into a single composite outcome defined as the time to the first of a set of qualifying events. For example, in type 1 diabetes it is of interest to evaluate the effect of an intervention with respect to the time to

*The Biostatistics Center, The George Washington University, 6110 Executive Blvd., Rockville, MD 20852, U.S.A., ibebu@bsc.gwu.edu.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

onset or progression of microvascular complications affecting the eyes (retinopathy), kidneys (nephropathy) and/or nerves (neuropathy). As another example, many cardiovascular outcome trials use the time to MACE (major adverse cardiovascular event) as a composite outcome defined as the time to cardiovascular death or non-fatal myocardial infarction or non-fatal stroke, whichever occurs first. This approach reduces a set of event times to a single time, the minimum of the set. Typically, the composite outcome time-to-first-event (TTFE) is then analyzed using simple univariate survival analysis methods such as the log-rank test or a Cox PH model. However, a common concern is that a TTFE analysis fails to capture the entire burden of disease by using only the first individual event and ignoring any subsequent events within the same subject. There is also a potential loss of power owing to analysis of only the first of a possible set of observed events.

Alternately, Lachin and Bebu [1] describe the application of other tests of the joint marginal null hypothesis including a T^2 -like test against a global alternative, an "average test" against the alternative of a common group difference for all outcomes and the Wei-Lin-Weissfeld robust global test. Herein, we focus on the application of the Wei-Lachin test of the joint marginal null hypothesis versus a one-directional alternative. This test is based on a separate analysis of the time to each component event, and thus employs the times to all component events, not just the first. The test is based on the average of the Cox PH model coefficients for the component events, and an average hazard ratio over all events and confidence limits are provided. Lachin and Bebu [1] also show that this test can have greater power than the composite analysis of the time to the first event.

Another common criticism of the composite outcome analysis is that all individual outcomes are weighted equally, although clearly not all have the same importance to patients (e.g., non-fatal MI versus cardiovascular death) [2]. To address this drawback it has been proposed that the groups be compared with respect to the individual outcomes when ordered from the most to least severe [3, 4]. Two subjects, one from each group, are first compared based on the most severe outcome, and the subject with a more preferable outcome (if either) is declared a winner. If the comparison based on the most severe outcome is inconclusive (e.g., when neither one of the two participants experienced that event type), then the subjects are compared based on the second most severe outcome, and the procedure is continued until a winner is declared or the comparison is inconclusive. The win ratio (WR) is defined as the ratio of the probability that a subject in the first group is a winner to the probability that a subject in the second group is a winner, and WR values larger than 1 are evidence that the first treatment is preferable to the second one. Further, several statistical tests and confidence intervals for a win ratio analysis have been proposed [3, 4, 5, 6, 7, 8].

However, the relationship of the joint distributions of the component events with the operating characteristics of the composite analysis, and those of the other analyses of the multiple events, have not been explored. Our goal is to assess the ability of these methods to capture the treatment group differences in the total burden of disease, as represented by the differences in the joint marginal distributions of the individual outcomes.

Herein, the operating characteristics of these analyses of multiple event times are illustrated using a bivariate exponential distribution based on a shared frailty model [1, 10]. These are

further described using simulations and an analysis of data from the Diabetes Control and Complications Trial (DCCT). We show that there are circumstances where the composite TTFE analysis can provide a biased test of the hypothesis of joint marginal equality of the survivor functions, but that the Wei-Lachin test is not affected. The presentation focuses on the case when all outcomes are observable up to independent censoring. The case of semi-competing risks when one event censors the others is discussed in Section 7.

2. Hypotheses to be Tested

Consider the case of two groups ($i = 1, 2$) with two outcomes with event times (T_{ia}, T_{ib}) in group i and bivariate survival functions $S_i(t_a, t_b) = P(T_{ia} > t_a, T_{ib} > t_b)$, $i = 1, 2$. Let ψ_j denote the hazard ratio for event j , $j = a, b$. Further, let S_{ia} and S_{ib} denote the marginal distributions within the i th group for each outcome.

The goal is to compare the two groups with respect to the entire burden of disease as captured by the two event times T_a and T_b . While event times can be right censored in practice, herein we assume no censoring, unless otherwise indicated. Thus, the objective is to test the joint null hypothesis

$$H_{all(0)}: S_1(t_a, t_b) = S_2(t_a, t_b) \quad \text{for all } t_a \geq 0 \text{ and } t_b \geq 0. \quad (1)$$

Unfortunately there is no convenient test of $H_{all(0)}$, such as a Kolmogorov-Smirnov test, especially one that allows for bivariate right censored times.

One approach could then be to test the hypothesis $H_{m(0)}$ of equality of the marginal hazards between groups for each event

$$H_{m(0)}: S_{1a}(t_a) = S_{2a}(t_a) \text{ for all } t_a \geq 0 \text{ and } S_{1b}(t_b) = S_{2b}(t_b) \text{ for all } t_b \geq 0, \quad (2)$$

which provides a complete description of the burden of disease. It is clear that $H_{all(0)} \Rightarrow H_{m(0)}$ and, rejection of $H_{m(0)}$ implies rejection of $H_{all(0)}$. Thus, an α -level test of $H_{m(0)}$ provides an α level test of $H_{all(0)}$.

Instead, the most widely used approach is to first define the composite time-to-first-event outcome in the i th group as $T_{ic} = \min(T_{ia}, T_{ib})$, with survival function $S_{ic}(t_c) = P(T_{ic} > t_c) = S_i(t_c, t_c)$, and to then test the null hypothesis

$$H_{c(0)}: S_1(t, t) = S_2(t, t). \quad (3)$$

This is equivalent to the null hypothesis of equality of all-cause hazards (the sum of the cause-specific hazards) between the two groups. A simple (1 df) logrank test of the composite outcome defined as the time to the first event provides an unbiased test for (3). However, it does not provide an unbiased test of the hypothesis of joint marginal equality

$H_{m(0)}$. In fact neither the hypothesis of marginal equality, nor the overall joint null hypothesis, are nested within the composite test null hypothesis. The basic feature of the composite analysis is that it is sensitive to a difference in correlations between groups whereas other methods such as the Wei-Lachin analysis are not.

Let ψ_a and ψ_b denote the hazard ratios (groups 1:2 herein) comparing the two groups for each outcome and let ψ_c denote that for the composite outcome. Then, the hypothesis of marginal equality $H_{m(0)}$ implies the hypothesis of joint null hazard ratios $H_{h(0)}$: $\psi_a = \psi_b = 1$. Then an unbiased α -level test of $H_{h(0)}$ provides a like test of the hypothesis of marginal equality $H_{m(0)}$, and rejection of $H_{h(0)}$ implies rejection of $H_{m(0)}$. For example, an α -level test of $H_{h(0)}$ can be provided by two separate tests of the hypotheses $\psi_a = 1$ and $\psi_b = 1$ using a logrank test (or a Wald test for the treatment group hazard ratio in a Cox model) with a Bonferroni adjustment for 2 tests, or with greater power accounting for the correlation between the two tests [11]. Such tests are unbiased because they are based only on the marginal distributions of the component events.

However, a difference in correlations ($\rho_{1ab} - \rho_{2ab}$) can result in a meaningful difference in the joint distributions. With no censoring, a correlation of 1 means that each subject experiences both events at the same time so that only 1 event time is observed for each subject. Conversely a correlation of 0 means that each subject will experience separate event times for each outcome. Thus, the treatment group with the smaller correlation will tend to have more events so that the times to the first event will be smaller in that group than the other, even in the presence of joint marginal equality.

Thus, it is possible for the marginal distributions to be equal among groups but for the joint distributions to differ between groups ($S_1 - S_2$) owing to a group difference in the correlations among the component event times. In this case a composite analysis TTFE test of the hypothesis $S_{1c} = S_{2c}$ will have a probability of rejection $> \alpha$ owing to the difference in correlations, and thus provides a biased test of $H_{m(0)}$.

Likewise, it is possible for the marginal distributions to differ between groups but for there to be little difference in the composite outcome distributions leading to a low probability of significance for the TTFE test that again provides a biased test of $H_{m(0)}$.

3. Properties of the Time to the First Event Analysis

The probabilities of significance of the different tests under the different hypotheses are illustrated using a simple bivariate exponential model. Herein we only consider two-sided tests, but the same conclusions also apply to one sided tests [1].

Let $\text{Expon}(\lambda)$ denote the exponential distribution with mean $1/\lambda$ and variance $1/\lambda^2$ ($\lambda > 0$). Also let $T_{11} \sim \text{Expon}(\lambda_{1a} - \lambda_{1F})$, $T_{12} \sim \text{Expon}(\lambda_{1b} - \lambda_{1F})$, and $T_{23} \sim \text{Expon}(\lambda_{2F})$, all independent exponentially distributed random variables with $\lambda_{ij} > \lambda_{iF} > 0$, $i = 1, 2$, $j = a, b$. In this model, T_{23} represents a shared frailty that induces a correlation among the event times. It follows that

$$\begin{aligned} T_{ia} &= \min(T_{i1}, T_{i3}) \sim \text{Expon}(\lambda_{ia}) \\ T_{ib} &= \min(T_{i2}, T_{i3}) \sim \text{Expon}(\lambda_{ib}), \end{aligned} \quad (4)$$

and the two times are correlated [1, 10]. We show in Appendix A (Eq. (34)) that this frailty model yields a joint survival distribution equal to that of Marshall and Olkin [12].

Then let

$$\psi_j = \frac{\lambda_{1j}}{\lambda_{2j}} \quad (5)$$

denote the hazard ratio comparing groups for the j th outcome. We are then interested in testing the joint null marginal hypothesis

$$H_{m(0)}: [\psi_a \psi_b] = [1 \ 1] \text{ or } \lambda_{1a} = \lambda_{2a} (= \lambda_a) \text{ and } \lambda_{1b} = \lambda_{2b} (= \lambda_b), \quad (6)$$

against the alternative

$$H_{m(1)}: [\psi_a \psi_b] \neq [1 \ 1] \text{ or } [\lambda_{1a} \ \lambda_{1b}] \neq [\lambda_{2a} \ \lambda_{2b}]. \quad (7)$$

This joint marginal null hypothesis could be tested against the omnibus alternative, for example, using a T^2 -like 2 df test. Alternately, one might wish to test the joint marginal null hypothesis $H_{m(0)}$ against a one-directional alternative specified as

$$\tilde{H}_{m(1)}: [\psi_a \psi_b] \geq [1 \ 1] \text{ and } \min(\psi_a, \psi_b) > 1. \quad (8)$$

This can be tested by the Wei-Lachin 1 df test that is based on the simple sum of the log hazard ratios [1].

Conversely, the composite analysis is based on the time to the first event encountered, which in the i th group is defined as

$$\begin{aligned} T_{ic} &= \min(T_{ia}, T_{ib}) = \min(T_{i1}, T_{i2}, T_{i3}) \\ &\sim \text{Expon}(\lambda_{ic}), i = 1, 2, \end{aligned} \quad (9)$$

with hazard rate

$$\lambda_{ic} = \lambda_{ia} + \lambda_{ib} - \lambda_{iF}. \quad (10)$$

The resulting hazard ratio for the composite event is

$$\psi_c = \frac{\lambda_{1a} + \lambda_{1b} - \lambda_{1F}}{\lambda_{2a} + \lambda_{2b} - \lambda_{2F}}. \quad (11)$$

We then wish to test the null versus alternative hypotheses

$$H_{c(0)}: \psi_c = 1 \text{ versus } H_{c(1)}: \psi_c \neq 1. \quad (12)$$

Under this exponential frailty model, without censoring, a simple expression for the correlation between T_a and T_b (or equivalently $1/\hat{\lambda}_a$ and $1/\hat{\lambda}_b$) within each group can easily be obtained. Suppressing the group subscript (i), within a group,

$$\begin{aligned} \text{Cov}(T_a, T_b) &= E(T_a \cdot T_b) - E(T_a) \cdot E(T_b) \\ &= E(T_a \cdot T_b) - 1/\lambda_a \cdot 1/\lambda_b, \end{aligned} \quad (13)$$

and

$$\text{Corr}(T_a, T_b) = \lambda_a \lambda_b \cdot \text{Cov}(T_a, T_b). \quad (14)$$

Conditioning on T_3 one obtains

$$\begin{aligned} E(T_a \cdot T_b) &= E(\min\{T_1, T_3\} \cdot \min\{T_2, T_3\}) \\ &= E(E(\min\{T_1, T_3\} | T_3) \cdot E(\min\{T_2, T_3\} | T_3)) \\ &= E\left(\frac{1 - \exp\{-\lambda_1 T_3\}}{\lambda_1} \cdot \frac{1 - \exp\{-\lambda_2 T_3\}}{\lambda_2}\right) \\ &= \frac{\lambda_F}{\lambda_1 \lambda_2} \left(\frac{1}{\lambda_F} - \frac{1}{\lambda_1 + \lambda_F} - \frac{1}{\lambda_2 + \lambda_F} + \frac{1}{\lambda_1 + \lambda_2 + \lambda_F}\right) \\ &= \frac{\lambda_F}{(\lambda_a - \lambda_F)(\lambda_b - \lambda_F)} \left(\frac{1}{\lambda_F} - \frac{1}{\lambda_a} - \frac{1}{\lambda_b} + \frac{1}{\lambda_a + \lambda_b - \lambda_F}\right). \end{aligned} \quad (15)$$

(Replacing (13) and (15) in (14) yields

$$\text{Corr}(T_a, T_b) = \frac{\lambda_a \lambda_b \lambda_F}{(\lambda_a - \lambda_F)(\lambda_b - \lambda_F)} \left(\frac{1}{\lambda_F} - \frac{1}{\lambda_a} - \frac{1}{\lambda_b} + \frac{1}{\lambda_a + \lambda_b - \lambda_F}\right) - 1. \quad (16)$$

This correlation coefficient is an increasing function in the frailty parameter λ_F and it can be estimated using the estimates for individual λ 's. Section 6 presents an example and describes estimation of the event hazards and the frailty hazards in the presence of censoring. These can then be substituted into the above expression to compute the estimated correlation.

Under the joint marginal null hypothesis $H_{m(0)}$ of equality of the marginal hazard functions among groups for each outcome, the composite outcome event hazard ratio (11) is simply

$$\psi_{c(0)} = \frac{\lambda_a + \lambda_b - \lambda_{1F}}{\lambda_a + \lambda_b - \lambda_{2F}}. \quad (17)$$

This composite event hazard ratio will differ from 1 (i.e. be biased) when the frailty hazard differs between the two groups ($\lambda_{1F} \neq \lambda_{2F}$) and/or $\text{Corr}(T_a, T_b) \neq 0$, and from (17), the group with the smaller frailty hazard, and lesser correlation among the event times, will have the higher composite event hazard rate, and thus will experience more composite outcome events than will the group with the higher frailty hazard and correlation. In this case a test of the hypothesis $\psi_{c(0)} = 1$ will reach statistical significance with a much higher probability than will an α -level test of the hypothesis of joint marginal equality $H_{m(0)}$.

It also follows from (11) that combinations of hazards will exist such that $\lambda_{1a} + \lambda_{1b} - \lambda_{1F} = \lambda_{2a} + \lambda_{2b} - \lambda_{2F}$ so that the marginal hazards differ and the correlations (frailties) also differ in a manner such that the composite hazard ratio approaches 1; that is, the difference in correlations cancels the difference in hazards. In that case a test of the hypothesis $\psi_{c(0)} = 1$ will fail to reach statistical significance with a much higher probability than will an unbiased test of the hypothesis of joint marginal equality $H_{m(0)}$. More generally, combinations of hazards can lead to a composite hazard ratio above or below 1.

4. Properties of the Win Ratio and Win Difference Measures

We likewise show that an analysis using the win ratio or win difference may also be biased with respect to $H_{m(0)}$. We first introduce some notation for an individual outcome, and then move to the composite outcome.

Following Bebu and Lachin [7], for two subjects χ_1 and χ_2 , one from each group, for a given outcome (a or b), let (T_1, C_1) and (T_2, C_2) denote the time of the event and time of censoring (only one of which is observed). Then χ_1 has a more favorable outcome than χ_2 , denoted by $\chi_1 \succ \chi_2$, if

$$\min\{C_1, C_2, T_1\} > T_2. \quad (18)$$

Equivalently, χ_2 may have a less favorable outcome than χ_1 , denoted by $\chi_2 < \chi_1$. Similarly, the comparison is non-informative if

$$\min\{C_1, C_2\} < \min\{T_1, T_2\}. \quad (19)$$

The individual outcomes are ordered from the most to least severe. Two subjects are compared based on the most severe outcome, and the subject with the more favorable outcome of that type (Eq. (18)), if any, is declared a winner. If the comparison is inconclusive (i.e., (19) holds), then the two subjects are compared with respect to the second most severe outcome, and the procedure is continued until a winner is declared or the comparison is inconclusive after examining all possible outcomes.

With two individual outcomes (a, b) , the probability that the subject from the first group (χ_1) is a winner is the probability that χ_1 has a more favorable first outcome (a) , plus the probability that the comparison based on the first outcome is inconclusive but χ_1 has a more favorable second outcome (b) than the subject from the second group.

We now derive this probability for the exponential frailty model introduced in Section 3, and we further consider random exponential censoring with hazard rate τ_i , $i = 1, 2$. One has

$$\begin{aligned} P(\chi_1 > \chi_2) &= P(\min\{C_1, C_2, T_{1a}\} > T_{2a}) \\ &+ P(\min\{C_1, C_2\} < \min\{T_{1a}, T_{2a}\} \& \min\{C_1, C_2, T_{1b}\} > T_{2b}). \end{aligned} \quad (20)$$

Under the assumption of joint independent censoring and independent groups, then $(C_1, C_2) \perp (T_{1a}, T_{2a})$ and $[C_1, C_2, T_{1a}, T_{2a}]$ are jointly independent exponential random variables. Then, the first term in (20) is given by

$$P\left(\min\{C_1, C_2, T_{1a}\} > T_{2a}\right) = \frac{\lambda_{2a}}{\tau + \lambda_{1a} + \lambda_{2a}}, \quad (21)$$

where $\tau = \tau_1 + \tau_2$. Using (4), the second term in (20) can be rewritten as

$$P(C < \min\{T_{11}, T_{13}, T_{21}, T_{23}\} \& \min\{C, T_{12}, T_{13}\} > \min\{T_{22}, T_{23}\}), \quad (22)$$

where $C = \min\{C_1, C_2\}$. This expression can be further simplified as follows. First, $C < T_{13}$ implies $\min\{C, T_{13}\} = C$, and therefore $\min\{C, T_{12}, T_{13}\} = \min\{C, T_{12}\}$. Second, $C < T_{23}$ implies $T_{12} > \min\{T_{22}, T_{23}\}$. It follows that (22) simplifies to

$$P(C < \min\{T_{11}, T_{13}, T_{21}, T_{23}\} \& \min\{C, T_{12}\} > T_{22}). \quad (23)$$

Since the T 's in the expression above are independent, it follows that, conditional on censoring C , the two events $C < \min\{T_{11}, T_{13}, T_{21}, T_{23}\}$ and $\min\{C, T_{12}\} > T_{22}$ are

independent, and therefore the probability of the intersection is the product of the probabilities. Straightforward calculations then show that (23) is given by:

$$\tau \cdot \frac{\lambda_{2b} - \lambda_{2F}}{\lambda_{1b} + \lambda_{2b} - \lambda_{1F} - \lambda_{2F}} \cdot \left(\frac{1}{\tau + \lambda_{1a} + \lambda_{2a}} - \frac{1}{\tau + \lambda_{1a} + \lambda_{2a} + \lambda_{2b} - \lambda_{2F}} \right). \quad (24)$$

The probability that the subject in the first group is a winner (i.e., $P(\chi_1 > \chi_2)$) is the sum of (21) and (24). Likewise calculations yield $P(\chi_1 < \chi_2)$. The win ratio and the win difference depend on the λ_{1F} and λ_{2F} parameter values even under the joint marginal null hypothesis ($\lambda_{1a} = \lambda_{2a}$, $\lambda_{1b} = \lambda_{2b}$), and therefore a win ratio/difference analysis does not provide an α -level test for the joint marginal null hypothesis $H_{m(0)}$.

5. Simulation

Several methods are compared for testing the joint marginal null hypothesis of null hazard ratios

$$H_{h(0)}: \psi_a = \psi_b = 1,$$

in the context of the shared frailty exponential model introduced in Section 3.

For this model, the joint null hypothesis $H_{all(0)}$ is the intersection of the null hypotheses of joint marginal equality $H_{m(0)}$ and equality of frailties $H_{f(0)}$, or $H_{all(0)} = H_{m(0)} \cap H_{f(0)}$, where

$$\begin{aligned} H_{m(0)} &: \lambda_{1a} = \lambda_{2a} \text{ and } \lambda_{1b} = \lambda_{2b}; \text{ and} \\ H_{f(0)} &: \lambda_{1F} = \lambda_{2F}. \end{aligned} \quad (25)$$

where λ_{iF} is the frailty term in the i th group.

The results of 10,000 simulations are reported in Table 1 for various combinations of parameter values using $n = 100$ and $\alpha = 0.05$. For each model the applicable null and/or alternative hypotheses are shown.

The first method is the logrank test for the time to the first event (denoted by TTFE), while the second is the Bonferroni adjustment for the logrank test applied to the two marginals. A T^2 type test (T^2) and the Wei-Lachin test are based on the joint distribution of $(\hat{\psi}_a, \hat{\psi}_b)$ using a sandwich estimate of the variance-covariance matrix obtained after fitting separate Cox PH models for the two outcomes a and b [13].

The first nine simulation models correspond to the null joint marginal hypothesis, $H_{m(0)}$: $\lambda_{1a} = \lambda_{2a}$ and $\lambda_{1b} = \lambda_{2b}$, and therefore the results represent empirical type I error probabilities under this hypothesis. Models 1 and 5 also have the same frailty hazard ($\lambda_{1F} = \lambda_{2F}$) so that the hypothesis of equal joint bivariate survival distributions $H_{All(0)}$ applies, and all tests provide a rejection probability ~ 0.05 . The other models (2–4 and 6–9) have jointly equal

marginal hazards but have unequal frailty hazards ($\lambda_{1F} \neq \lambda_{2F}$) so that $H_{m(0)}$ is true but $H_{f(0)}$ is not. Since $\rho_1 > \rho_2$, the difference favors group 1. In these cases, as expected from Section 3, the composite analysis based on the time to the first event (TTFE) has increasing probability of significance as the difference in frailties increases, reaching 0.71 for frailty parameters $\lambda_{1F} = 0.2$ versus $\lambda_{2F} = 0.025$, thus providing a biased test of the joint marginal null hypothesis $H_{m(0)}$. However, in these models the other three approaches provide α -level tests for the joint marginal null hypothesis $H_{m(0)}$. Not shown, a test based on the win-ratio has properties similar to those of the TTFE.

Simulation 10 presents results under the alternative hypothesis that the marginal hazards differ for both individual outcomes favoring group 2, but there is no difference in frailties. Therefore, the estimated probabilities correspond to the power of each test to reject the joint marginal null hypothesis. In this case the TTFE test is unbiased but it nevertheless has less power than the other marginal tests.

The results of simulations with different marginal hazards and different frailty parameters are presented in models 11–13. As expected from Eq. (11), the TTFE test has greater power than the other methods when the frailty parameter is lower in the group with higher total marginal hazards (simulation 11), and generally lower power otherwise (simulations 12 and 13).

Finally, simulation 14 assumes that group 2 has a lower hazard for the two outcomes (fewer events) but it also has a lower correlation (more events), the effects potentially canceling. The parameters were chosen such that the composite time to the first event is the same in the two groups (i.e., $\psi_c = 1$, see Eq. (11)). As expected, the power of the composite analysis is approximately 5%, while the other approaches provide considerable power to reject the null hypothesis of marginal equality.

Table 1 also presents simulations using the same parameters (hazards) in the presence of right censoring. A total $n = 200$ (100 per group) is employed assuming administrative censoring of all subjects after $L = 3$ years of follow-up. The analysis based on the time to the first event (TTFE) again has increasing probability of significance as the difference in frailties increases, but not as great as with no censoring, reaching 0.595 versus 0.705 in model 9, substantial inflation nonetheless.

6. Example

The above simulation demonstrates the theoretical concern that a difference between treatment groups in the event-time correlations can increase or decrease the probability that the TTFE test will be significant, even when there is no difference in the marginal hazards. The next question is whether such a treatment group difference is of practical concern.

Under model (4), allowing for random right censoring of some event times, the MLE's for the parameters are given by

$$\hat{\lambda}_{ij} = \frac{D_{ij}}{X_{ij}}, \quad \hat{\lambda}_{ic} = \frac{D_{ic}}{X_{ic}}, \quad \hat{\lambda}_{iF} = \hat{\lambda}_{ia} + \hat{\lambda}_{ib} - \hat{\lambda}_{ic}, \quad (26)$$

where D_{ij} denotes the number of observed events among subjects, and X_{ij} the total follow-up time to the event or right censoring, of type $j = a, b$ in group $i = 1, 2$. Similarly, D_{ic} and X_{ic} are defined for the time to the first event in group $i = 1, 2$. The numbers of events in the combined groups are designated as D_a , D_b and D_c as are the corresponding hazards $\hat{\lambda}_a$, $\hat{\lambda}_b$ and $\hat{\lambda}_c$. The number of subjects with both events is designated as D_{iab} and D_{ab} within the i th group and total.

The common hazard rates in both groups under the null hypothesis (λ_a and λ_b) can be estimated using data from both groups by

$$\hat{\lambda}_j = \frac{D_{1j} + D_{2j}}{X_{1j} + X_{2j}} \quad (27)$$

for $j = a, b$. From (26) the frailty hazard rates within each group, assuming the joint marginal null hypothesis, are obtained as $\hat{\lambda}_{iF} = \hat{\lambda}_a + \hat{\lambda}_b - \hat{\lambda}_{ic}$ from which an estimate of the hazard ratio under the joint null in (17) ($\hat{\psi}_{c(0)}$) is obtained.

To illustrate the potential consequences of different correlations between individual outcomes in an actual study, consider data from the DCCT study [14]. Briefly, 1441 subjects with type 1 diabetes were randomized to receive either standard therapy or intensive therapy, denoted herein by groups 1 and 2, respectively. Two outcomes are considered, one related to retinopathy and the other one to nephropathy. The retinopathy outcome is a 3-step progression on the ETDRS scale, while the nephropathy outcome is microalbuminuria, defined as albumin excretion rate (AER) ≥ 30 mg per 24 hours [14].

With a mean follow-up of 6.5 years since randomization, there were $D_a = 170$ 3-step progression events, and $D_b = 475$ microalbuminuria events, $D_c = 546$ participants had at least one of them, and $D_{ab} = 99$ had both [14]. The following describes the estimates of the hazard rates and correlation within each group, plus those in the two groups combined under the joint null hypothesis $H_{A|K(0)}$.

Group (<i>i</i>)	Microalbuminuria		3-step progression		Frailty	Joint	Composite		Corr(T_a , T_b)
	D_{ia}	$\hat{\lambda}_{ia}$	D_{ib}	$\hat{\lambda}_{ib}$	$\hat{\lambda}_{iF}$	D_{iab}	D_{ic}	$\hat{\lambda}_{ic}$	
Standard (1)	112	0.0287	307	0.0888	0.0121	74	345	0.1055	0.1144
Experimental (2)	58	0.0147	168	0.0467	0.0028	25	201	0.0587	0.0471
Combined Joint Null	170	0.0217	475	0.0673	0.0075	99	546	0.0815	0.0920

where the "Joint" column refers to the number of subjects experiencing both events.

A composite TTFE Cox PH model analysis of the 546 composite outcome events yields a hazard ratio estimate of 1.940 with a 2-sided p -value of $1.27e - 13$, highly significant.

Alternately, under a simple univariate exponential model, the empirical estimate of the composite outcome hazard ratio is $0.1055/0.0587 = 1.797$ and a simple test of the log(HR) under this model yields $Z = \ln(1.797)\sqrt{546}/2 = 6.85$ with 2-sided $p = 1.84e - 12$. In other words, the standard therapy is associated with about a 80% higher risk ($HR \approx 1.8$) of a composite event relative to the intensive therapy under the observed frailties in the two groups. These exponential model results are similar to those of the Cox model, with a somewhat smaller estimate of the composite hazard ratio in the exponential model.

Since $\hat{\lambda}_{1F} > \hat{\lambda}_{2F}$ there is a higher correlation among the event times in group 1 than in group 2. Under the joint marginal null hypothesis, from the expression for $\psi_{c(0)}$ in (17), the group with the smaller frailty hazard (group 2 in this case) would be expected to have more composite events under the marginal null hypothesis. However, in this example the marginal hazards for both events are higher in standard group 1, so that group 1 has the higher composite event hazard rate and more composite events observed even though it has the larger correlation.

This implies that the group difference in the composite outcome hazards would have been even greater had there not been a difference in frailties. To address this, the frailty in the combined groups can be estimated as simply $\hat{\lambda}_F = (\hat{\lambda}_{1F} + \hat{\lambda}_{2F})/2 = 0.0074$. When substituted into (11) for λ_{1F} and λ_{2F} this yields $\hat{\psi}_c = 2.038$. Thus, there is about a 103% higher risk in the standard group (1) if the frailty parameters are assumed equal in the two groups, greater than the 94% increase observed in the Cox model or the 80% increase in the bivariate exponential model with a difference in correlations. In this case the difference in frailties dampens the hazard ratio of the composite outcome.

This in turn implies that the difference in frailties would have the expected effect on risk had there been no difference in the component hazards. The last row of the above table shows the estimates of the hazards of each outcome under the joint marginal null hypothesis in the two groups combined. Evaluating the expression for $\psi_{c(0)}$ in (17) using the estimated common hazards $\hat{\lambda}_a = 0.0217$ and $\hat{\lambda}_b = 0.0673$ for the two events, but with the group-specific estimates of the frailties $\hat{\lambda}_{iF}$ ($i = 1, 2$), yields a hazard ratio estimate $\hat{\psi}_{c(0)} = 0.8922$ under the joint

marginal null hypothesis rather than the null value of 1. Thus, had there not been a difference in hazards between groups, the standard group (1) with the higher correlation would have had a 10% lower risk as would be expected. These additional computations illustrate the potential impact of the difference in frailties between groups, and the difference in correlations of the component events, on the composite outcome event rates.

For a win-ratio analysis, using Eq. (20), the probability of a win comparing intensive versus standard therapy is $P(\chi_2 > \chi_1) = 0.2126$, while that comparing standard versus intensive therapy is $P(\chi_1 > \chi_2) = 0.0918$, for a win ratio of 2.316. Now assume equal frailties in the two groups. Replacing λ_{1F} and λ_{2F} with $\hat{\lambda}_F$ in (21) and (24) yields 0.2252 for intensive versus standard therapy and 0.0841 for comparing standard versus intensive, for a win ratio of 2.678. In other words, there is an approximately 2.3 fold higher chance of a favorable outcome receiving the intensive therapy than the standard therapy using the observed frailties, and a 2.7 fold higher chance assuming equal frailty parameters in the two groups. This is consistent with the results obtained above using the time to the first event composite outcome.

Alternatively, a T^2 test or a Wei-Lachin analysis can be employed (Appendix B). Using separate Cox models for each event type, the treatment log hazard ratios for microalbuminuria and 3-step progression are 0.6788 and 0.7433, respectively, while a sandwich estimate of the variance-covariance matrix is given by

$$\begin{pmatrix} 0.02643 & 0.00173 \\ 0.00173 & 0.00979 \end{pmatrix}.$$

The T^2 omnibus test yields value $\chi^2_{(2df)} = 67.88$ ($p = 1.77e - 15$), while the Wei-Lachin test yields $z = 7.18$ ($p = 8.5e - 14$), both highly significant and with smaller p-values than provided by the composite Cox PH model above ($p = 1.27e - 13$).

Also note that the time-to-first-event analysis only employed the 546 first event times, while the T^2 test and the Wei-Lachin analysis used all 645 observed events, including both the microalbuminuria and the 3-step progression events for the 99 participants who reached both. This may explain the slightly lower p-values for the T^2 test and the Wei-Lachin analysis compared to the composite outcome analysis.

7. Semi-Competing Risks

Additional issues arise in the context of competing risks. We illustrate this for semi-competing risks [9], when one of the individual outcomes (such as death) censors the follow-up times for other individual outcomes. It can be shown that the composite time to the first event is unaffected by the competing risk mechanism, and we also show that neither is the win ratio analysis.

To see this, assume a is an absorbing competing risk event, while the event b is not absorbing. Let T_a denote the time to event a , and T_b the latent time to event b assuming no

censoring by the absorbing event a . Since event a censors event b , T_b is only observable if $T_b < T_a$. It follows that, with censoring C , the observed composite time to the first event is $\tilde{T}_c = \min\{T_a, \min\{T_a, T_b\}, C\} = \min\{T_a, T_b, C\}$, the same as in the case when event a does not censor the event b .

We now consider the win ratio analysis with semi-competing risks. With C_i denoting the censoring time in group i , since the event b is only observed if $T_{ib} < \min\{C_i, T_{ia}\}$, it follows that $C_{ib} = \min\{C_i, T_{ia}\}$ is the censoring mechanism for T_{ib} , $i = 1, 2$. Then Eq. (20) becomes

$$P(\chi_1 > \chi_2) = P(\min\{C_1, C_2, T_{1a}\} > T_{2a}) + P(\min\{C_1, C_2\} < \min\{T_{1a}, T_{2a}\} \& \min\{C_{1b}, C_{2b}, T_{1b}\} > T_{2b}). \quad (28)$$

The first term on the right hand side is the same as in (20), while the second term can be rewritten as

$$P(\min\{C_1, C_2\} < \min\{T_{1a}, T_{2a}\} \& \min\{C_1, T_{1a}, C_2, T_{2a}, T_{1b}\} > T_{2b}),$$

where $\min\{C_1, C_2\} < \min\{T_{1a}, T_{2a}\}$ implies $\min\{C_1, T_{1a}, C_2, T_{2a}, T_{1b}\} = \min\{C_1, C_2, T_{1b}\}$. It follows that (28) is equivalent to (20). Notice that this equivalence will not hold if a less severe individual outcome censors a more severe one.

Let us now consider a marginal analysis. With semi-competing risks, the marginal hazards function for the absorbing event a (i.e., death)

$$\lambda_a(t) = \lim_{\Delta t \searrow 0} \frac{P(t \leq T_a < t + \Delta t | T_a \geq t)}{\Delta t}$$

is estimable without additional untestable assumptions, as is the conditional hazard function of the non-absorbing time T_b given $T_a > t$,

$$\lambda_{b|a}(s|t) = \lim_{\Delta s \searrow 0} \frac{P(s \leq T_b < s + \Delta s | T_a \geq s, T_b \geq t)}{\Delta s}, 0 \leq s \leq t.$$

These two hazards uniquely determine the survival function $S_{a,b}(t_a, t_b)$ in the lower wedge corresponding to observable data (i.e., $0 \leq t_b \leq t_a$)

$$S_{a,b}(t_a, t_b) = \exp\left\{-\int_0^{t_a} \lambda_a(u) du - \int_0^{t_b} \lambda_{b|a}(u|t_a) du\right\}. \quad (29)$$

Consider now the case of two groups. Within each group, fitting separate risk models (such as Cox PH) to the absorbing and the non-absorbing outcomes yields estimates of $\lambda_{ia}(t)$ and $\lambda_{i,b|a}(t|t)$, the marginal hazards for the absorbing outcome a and the conditional hazards for

the non-absorbing outcome b , respectively ($i = 1, 2$). This then yields estimates for the hazard ratios ψ_a and $\psi_{b|a}$ comparing the two groups.

One can then test the null hypothesis

$$\tilde{H}_{m(0)}: \lambda_{1a}(t) = \lambda_{2a}(t) \text{ and } \lambda_{1,b|a}(t|t) = \lambda_{2,b|a}(t|t) \quad (30)$$

using a T^2 -like test or the Wei-Lachin test employing a sandwich estimate of the variance-covariance matrix obtained after fitting the two marginal (Cox) models.

The hazards function of the composite time-to-first event outcome is the sum of the marginal hazards for the absorbing event and the conditional hazards for the non-absorbing one,

$$\lambda_c(t) = \lambda_a(t) + \lambda_{b|a}(t|t), \quad (31)$$

which is the same as a Wei-Lachin analysis with equal weights for the two endpoints.

It follows that a composite time to the first event analysis is an α -level test for (30). However, there will be combinations of ψ_a and $\psi_{b|a}$ values under the alternative (i.e., different than 1) such that $\psi_c \approx 1$, and the composite test will have low power in these situations. As a simple illustration, assuming the survival function in group $i = 1, 2$ is

$$S_{i,a,b}(t_a, t_b) = \exp\{-\alpha_{ia} \cdot t_a - \alpha_{ib} \cdot t_b - \alpha_{iab} \cdot t_{ab}\},$$

with $\alpha_{ia}, \alpha_{ib}, \alpha_{iab} > 0$, then one obtains

$$\lambda_{ia}(t) = \alpha_{ia}, \quad \lambda_{i,b|a}(t) = \alpha_{ia} + \alpha_{iab} \cdot t \quad \text{and} \quad \lambda_{ic}(t) = \alpha_{ia} + \alpha_{ib} + \alpha_{iab} \cdot t.$$

Therefore, an improvement of $\delta > 0$ in the hazards for outcome a (i.e., $\alpha_{1a} - \alpha_{2a} = \delta$) and a similar reduction δ in the conditional hazards for b (i.e., $\alpha_{1,b|a} - \alpha_{2,b|a} = -\delta$) yields a hazard ratio of 1 for the composite analysis of time to the first event, but not for the T^2 -like test or the Wei-Lachin test with unequal weights for the two outcomes.

8. Discussion

Composite outcomes are common in medical applications, particularly in cardiovascular disease studies. They are usually employed to hopefully increase statistical power by increasing the incidence of composite events relative to that of an analysis of any one single component outcome, and also attempt to capture the joint treatment effect on several outcomes of interest. Examples include the analysis of the composite outcome of the time to the first event (TTFE) and the win ratio or difference for ordered outcomes.

Herein, we used a simple bivariate exponential model with shared frailties to show that the incidence of the composite outcome (or the fraction of "wins") can differ between two

groups even when, marginally, the hazard rates (survival functions) of the individual outcomes are the same in the two groups, and vice versa. The essential problem is that the composite TTFE test is sensitive to a difference between groups in the correlations among the event times whereas a test of the marginal differences is not. Further, the correlations may differ between groups when the marginal distributions are equal. Accordingly, the composite test provides an unbiased test of the null hypothesis of no difference between groups in the distribution of the composite time-to-first-event outcome, but it can also provide a biased test for the joint null hypothesis of equal marginal distributions.

Therefore, the composite TTFE outcome is appropriate and unbiased when the main interest is to compare treatments with respect to the time to the occurrence of the first event among those in the composite, irrespective of whether there are marginal group differences in the distributions of the component event times, and irrespective of whether there are group differences in correlations. However, the composite analysis provides a biased comparison of treatments with respect to differences in the marginal hazards of the individual outcomes, and thus fails to capture the treatment effects on the entire burden of disease.

While derivations herein were obtained in closed form for the bivariate frailty model considered, the same considerations apply more generally. For example, one can easily extend the frailty model to $K(K-2)$ outcomes by defining $T_j = \min\{\tilde{T}_j, T_F\}$, where $\tilde{T}_j \sim \text{Expon}(\lambda_j - \lambda_F)$ and $T_F \sim \text{Expon}(\lambda_F)$ are independent exponentially distributed times, $j = 1, \dots, K$. It follows that the TTFE composite outcome is again exponentially distributed with hazard rate $\sum_{j=1}^K \lambda_j - (K-1) \cdot \lambda_F$, which reduces to (10) for $K=2$.

We further used real data to show that such considerations may actually apply in clinical studies. In other words, differences in actual correlation structures between groups may occur that can lead to inflation of the probability of significance for the composite TTFE test even though there is no difference in the marginal rates.

We then considered the case of semi-competing risks, when one of the outcomes (such as death) censors an other (less severe) outcome. Censoring by that absorbing outcome (e.g., death) only appears if the absorbing event occurs *before* the non-absorbing one, and therefore the composite time to the first event analysis is not affected by having a competing risk. A similar argument applies to the win-ratio analysis when the absorbing outcome is more severe than the un-absorbing outcome. Since the absorbing event censors the non-absorbing one, only the marginal hazards for the absorbing event and the conditional hazards for the non-absorbing ones given no absorption are estimable without additional untestable assumptions. A T^2 -like test or a Wei-Lachin test can then be employed to test differences between two groups in terms of these two hazards. Since the hazards function for the composite outcome is the sum of the marginal hazards for the absorbing event and the conditional hazards for the non-absorbing event, it follows that the composite analysis tests the same parameter as the Wei-Lachin test with equal weights for the two outcomes and both provide an α level test for the null joint hypothesis of equal marginal hazards for the absorbing event and the conditional hazards for the non-absorbing event. However, there are combinations of these hazards under the alternative hypothesis which yield a composite

hazard ratio close to 1, in which cases the composite analysis will have very low power. That is not the case for the T^2 -like test.

It may be argued that reducing the time to the first event (the event-free time) among a set of individual outcomes represents a real benefit to the patient. However, the time to the first event does not capture the entire burden of disease, while the marginal hazards do a better job of describing all the individual outcomes because all observed events are counted.

Investigating the correlation between individual outcomes with censored data is a difficult problem in the presence of censoring [16]. However, several approaches have been proposed to account for such dependence, including marginal approaches [13], frailty models [16], and copula models [17]. Unlike the composite analysis, these methods provide unbiased statistical tests for the joint null hypothesis of equal marginal hazards, regardless of whether or not the correlations between individual outcomes are equal in the two groups. Careful consideration of the operating characteristics of various methods for analysis of composite outcome measures are in order before adopting them as primary outcomes in clinical studies.

A composite time-to-first-event analysis is appropriate when it is expected that the group (treatment) effect is in the same direction for all individual outcomes, but not necessarily of the same degree. However, the win ratio analysis and the Wei-Lachin approach can be tailored to cases where the individual outcomes are negatively correlated, such as when dealing with efficacy and adverse events outcomes.

As pointed out by one of the reviewers, additional considerations may apply. Using the DCCT example for illustration, since the marginal approaches will employ microalbuminuria events that occurred after a 3-step progression (and vice versa), the hazards of such subsequent events may be affected by exposure changes in response to the occurrence of the 3-step progression event. While conditional models similar to those for recurrent events [15] may help, further research is needed to fully describe such models.

In practice, one might wish to assess whether there is a difference in frailty hazards or correlations between groups, and if so, its effects on the operating characteristics of the composite analysis. Under the joint exponential model employed herein, as shown in the Example, one can estimate a common frailty for the outcomes within groups and on that basis compare the hazard ratio provided by the composite analysis versus an analysis with a common frailty.

Along the same lines one might then ask whether it is possible to test for a difference between groups in the frailties or correlations. It would be possible to obtain an estimate of the SE of each frailty hazard under the bivariate exponential distribution employed herein using the delta method or bootstrap on which basis a test could be constructed. Likewise, there are other methods for comparing the correlation matrices of event times between groups [16]. Such tests, however, will have very poor power.

Further, the bivariate exponential distribution assumptions may not apply in practice. Thus, it would be prudent to assess the exponential assumption before using this model to assess operating characteristics or conduct tests of significance or compute confidence limits. This

could be done visually by computing the simple Nelson-Aalen estimate of the hazard rate function for each type of event over time within each group. In doing so, however, note that it is sufficient for the hazard rates to be proportional between groups since the exponential model will apply approximately under a proportional hazards model assumption.

Thus, we conclude that alternative approaches that are not affected by treatment group differences in the correlations among outcomes, and which focus on the marginal hazard functions, are preferable in the analysis of multiple event times. These include the omnibus T^2 test on K df for K outcomes jointly, or separate tests of the K individual outcomes with a correction for multiple tests, or a 1 df one-directional test such as that of Wei and Lachin. Even when a composite time to the first event may be justifiable in terms of benefit to the patient, presenting marginal analyses for the multiple events will provide a better understanding of the entire burden of disease and illuminate potential heterogeneity in the effect of the intervention on these multiple outcomes.

Acknowledgment

The authors thank the DCCT/EDIC research group for permission to use the data presented in Section 5 (Example). The data employed are also available from the NIDDK Data Repository at <https://www.niddkrepository.org/> with specific reference to the data employed in the generation of reference (13) herein.

The authors are also grateful to the Editor, Associate Editor and the two referees for insightful and helpful comments that led to substantial improvements in the paper.

Funding

This work was partially supported by grant numbers U01-DK-094176 and U01-DK-094157 for the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study, and by grant U01-DK-098246 for the Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness (GRADE) Study, all from the National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK), NIH.

The first author (I.B.) was also supported by the Samuel W. Greenhouse Biostatistics Research Enhancement Award.

Appendix

9.

9. 9.1. Appendix A

The joint survival function under the exponential frailty model (4) is expressed as

$$\begin{aligned} P(T_a \geq t_a, T_b \geq t_b) &= P(\min\{T_1, T_3\} \geq t_a, \min\{T_2, T_3\} \geq t_b) \\ &= E(I(\min\{T_1, T_3\} \geq t_a, \min\{T_2, T_3\} \geq t_b | T_3)) \end{aligned} \quad (32)$$

where $I()$ is the indicator function, and the expectation in (32) is with respect to T_3 . Given T_3 , one has

$$P(\min\{T_1, T_3\} \geq t_a) = \begin{cases} 0 & \text{if } t_a \geq T_3 \\ \exp\{-\lambda_1 t_a\} & \text{if } t_a < T_3. \end{cases}$$

A similar result holds for $P(\min\{T_1, T_3\} \geq t_a)$, and therefore, conditional on T_3 ,

$$I(\min\{T_1, T_3\} \geq t_a, \min\{T_2, T_3\} \geq t_b) = \begin{cases} \exp\{-\lambda_1 t_a - \lambda_2 t_b\} & \text{if } \max\{t_a, t_b\} < T_3 \\ 0 & \text{otherwise.} \end{cases} \quad (33)$$

Replacing (33) in (32) yields

$$P(T_a \geq t_a, T_b \geq t_b) = \exp\{-\lambda_1 t_a - \lambda_2 t_b - \lambda_F \max\{t_a, t_b\}\}. \quad (34)$$

This is identical to the joint survival distribution of Marshall and Olkin (1967).

9. 9.2. Appendix B

Let β_j denote the log hazard ratio comparing the two groups for outcome j , so that $\beta_j = \log(\psi_j)$, $j = a, b$. Then the joint distribution of the estimated log hazard ratios for the individual outcomes is asymptotically normal

$$(\hat{\beta}_a, \hat{\beta}_b)' \sim N((\beta_a, \beta_b)', \Sigma),$$

where an estimate $\hat{\Sigma}$ of Σ can be obtained using a sandwich estimate after fitting separate Cox models for each of the individual outcomes. Hypothesis $H_{m(0)}$ is then equivalent to testing $\beta_a = \beta_b = 0$.

A T^2 -like test is given by

$$(\hat{\beta}_a, \hat{\beta}_b)' \cdot \hat{\Sigma}^{-1} \cdot (\hat{\beta}_a, \hat{\beta}_b),$$

which under $H_{m(0)}$ follows asymptotically a chi-squared distribution with 2 degrees of freedom.

The Wei-Lachin test is given by

$$\frac{\hat{\beta}_a + \hat{\beta}_b}{\sqrt{J' \hat{\Sigma} J}}, \quad J = (1, 1)',$$

which under $H_{m(0)}$ follows a standard normal distribution. Note that a Wei-Lachin analysis can also incorporate weights w_a and w_b for the two outcomes, which may potentially reflect different severities of the two outcomes a and b

$$\frac{W \cdot (\hat{\beta}_a, \hat{\beta}_b)'}{\sqrt{W \cdot \hat{\Sigma} \cdot W'}}, \quad W = (w_a, w_b),$$

which again follows asymptotically a standard normal distribution under $H_{m(0)}$.

References

1. Lachin JM, Bebu I. Application of the Wei-Lachin multivariate one-directional test to multiple event-time outcomes, *Clinical Trials* 2015; 12(6): 627–633. [PubMed: 26336199]
2. Stolker JM, Spertus JA, Cohen DJ, Jones PG, Jain KK, Bamberger E, Lonergan BB, Chan PS. Rethinking composite end points in clinical trials: insights from patients and trialists, *Circulation* 2014; 130: 1254–1261. [PubMed: 25200210]
3. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials, *Statistics in Medicine* 1999; 18: 1341–1354. [PubMed: 10399200]
4. Pocock SJ, Ariti CA, Collier TJ, Wand D. The win ratio: a new approach to the analysis of composite outcomes in clinical trials based on clinical priorities, *European Heart Journal* 2012; 33: 176–182. [PubMed: 21900289]
5. Buyse M Generalized pairwise comparisons of prioritized outcomes in the two-sample problem, *Statistics in Medicine* 2010; 29: 3245–3257. [PubMed: 21170918]
6. Luo X, Tian H, Mohanty S, Tsai WY. An alternative approach to confidence interval estimation for the win ratio statistics, *Biometrics* 2014; 71: 139–145. [PubMed: 25156540]
7. Bebu I, Lachin JM. Large Sample Inference for a Composite Analysis Based on Prioritized Components, *Biostatistics* 2016; 17: 178–187. [PubMed: 26353896]
8. Oakes D On the win-ratio statistic in clinical trials with multiple types of event, *Biometrika* 2016; 103(3): 742–745.
9. Fine JP, Jiang H, Chappell R. On semi-competing risks, *Biometrika* 2001; 88(4): 907–919.
10. Lachin JM. Applications of the Wei-Lachin multivariate one-sided test for multiple outcomes on possibly different scales. *PLoS ONE* 2014; 9(10): e108784. doi:10.1371/journal.pone.0108784 . Supplemental Information doi:10.1371/journal.pone.0108784.10.1371/journal.pone.0108784.s001. Supplemental Information doi: 10.1371/journal.pone.0108784.s001. [PubMed: 25329662]
11. Piper CB, Ritz C, Bisgaard H. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models, *Journal of the Royal Statistical Society, C* 2012; 61(2): 315–326.
12. Marshall AW, Olkin I. A multivariate exponential distribution, *Journal of the American Statistical Association* 1967; 62: 30–44.
13. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* 1989; 84: 1065–1073.
14. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes in the development and progression of long-term complications in insulin-dependent diabetes mellitus, *New England Journal of Medicine* 1993; 329: 977–986. [PubMed: 8366922]
15. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data, *Biometrika* 1981; 68: 373–389.
16. Hougaard P Analysis of multivariate data, New York: Springer-Verlag, 2000.
17. Shih JH, Louis TA (1995). Inferences on the association parameter in copula models for bivariate survival data, *Biometrics* 1995; 51: 1384–1399. [PubMed: 8589230]

Group (<i>i</i>)	Microalbuminuria		3-step progression		Frailty	Joint	Composite		
	D_{ia}	$\hat{\lambda}_{ia}$	D_{ib}	$\hat{\lambda}_{ib}$	$\hat{\lambda}_{iF}$	D_{iab}	D_{ic}	$\hat{\lambda}_{ic}$	$\text{Corr}(T_a, T_b)$
Standard (1)	112	0.0287	307	0.0888	0.0121	74	345	0.1055	0.1144
Experimental (2)	58	0.0147	168	0.0467	0.0028	25	201	0.0587	0.0471
Combined Joint Null	170	0.0217	475	0.0673	0.0075	99	546	0.0815	0.0920

Table 1.

Simulation results comparing the probability of reaching statistical significance at level $\alpha = 0.05$ under the exponential frailty model in Section 3 using the composite outcome test of time-to-first-event (TTFE), individual tests with the Bonferroni adjustment (B), a T^2 test of the joint marginal hypothesis (T^2), and the Wei-Lachin test employing a sandwich estimate of the variance-covariance matrix (WL). The null or alternative hypotheses reflected by the parameters in each model are also shown, indicated, respectively, by (0) or (1) subscript.

Model	H	$\lambda_{1a} = \lambda_{1b}$	λ_{2a}	λ_{2b}	λ_{2f}	λ_{2f}	ρ_1^*	ρ_2^*	No Censoring (n = 100)				With Censoring (n = 200)			
									TTFE	B	T^2	WL	TTFE	B	T^2	WL
1	$H_{A 00}$	0.2	0.20	0.2	0.10	0.100	0.333	0.333	0.0530	0.0526	0.0575	0.0553	0.0474	0.0481	0.0519	0.0495
2	$H_{m0>} H_{f(1)}$	0.2	0.20	0.2	0.10	0.075	0.333	0.231	0.0912	0.0500	0.0562	0.0570	0.0716	0.0455	0.0491	0.0520
3	$H_{m0>} H_{f(1)}$	0.2	0.20	0.2	0.10	0.050	0.333	0.143	0.1890	0.0505	0.0555	0.0532	0.1342	0.0463	0.0496	0.0474
4	$H_{m0>} H_{f(1)}$	0.2	0.20	0.2	0.10	0.025	0.333	0.067	0.3479	0.0511	0.0555	0.0523	0.2366	0.0449	0.0492	0.0496
5	$H_{A 00}$	0.3	0.30	0.3	0.20	0.200	0.500	0.500	0.0530	0.0480	0.0555	0.0531	0.0500	0.0451	0.0482	0.0505
6	$H_{m0>} H_{f(1)}$	0.3	0.30	0.3	0.20	0.100	0.500	0.200	0.3461	0.0509	0.0578	0.0519	0.2762	0.0427	0.0494	0.0482
7	$H_{m0>} H_{f(1)}$	0.3	0.30	0.3	0.20	0.075	0.500	0.143	0.4832	0.0497	0.0587	0.0571	0.3870	0.0483	0.0545	0.0541
8	$H_{m0>} H_{f(1)}$	0.3	0.30	0.3	0.20	0.050	0.500	0.091	0.6020	0.0481	0.0503	0.0499	0.4961	0.0468	0.0523	0.0494
9	$H_{m0>} H_{f(1)}$	0.3	0.30	0.3	0.20	0.025	0.500	0.043	0.7054	0.0490	0.0554	0.0535	0.5946	0.0471	0.0536	0.0526
10	$H_{m(1)>} H_{f(0)}$	0.3	0.25	0.2	0.10	0.100	0.200	0.286	0.7012	0.7469	0.7383	0.8106	0.5675	0.4544	0.4415	0.5347
11	$H_{m(1)>} H_{f(1)}$	0.3	0.25	0.25	0.10	0.200	0.200	0.667	0.9450	0.2708	0.2539	0.2477	0.8456	0.1623	0.1493	0.1609
12	$H_{m(1)>} H_{f(1)}$	0.3	0.25	0.2	0.25	0.150	0.714	0.500	0.1862	0.7252	0.7465	0.8125	0.1379	0.4333	0.4737	0.5434
13	$H_{m(1)>} H_{f(1)}$	0.3	0.25	0.25	0.25	0.100	0.714	0.250	0.1593	0.2697	0.2529	0.2506	0.1165	0.1602	0.1505	0.1652
14	$H_{m(1)>} H_{f(1)}$	0.3	0.30	0.2	0.20	0.100	0.500	0.250	0.0472	0.7324	0.7939	0.8161	0.0496	0.4427	0.5391	0.5460

* $\rho_i = \text{Corr}(T_{ia}, T_{ib})$, $i = 1, 2$, see Eq. (16).