Dealing with competing risks in clinical trials: How to choose the primary efficacy analysis?

James F. Troendle, Eric S. Leifer, Lauren Kunz

Office of Biostatistics Research, Division of Cardiovascular Sciences of the National Heart, Lung, and Blood Institute, NIH/DHHS, Bld RLC2 Room 9196, Bethesda, MD 20892, USA

Abstract

We investigate different primary efficacy analysis approaches for a 2-armed randomized clinical trial when interest is focused on a time to event primary outcome that is subject to a competing risk. We extend the work of Friedlin and Korn (2005) by considering estimation as well as testing and by simulating the primary and competing events' times from both a cause-specific hazards model as well as a joint subdistribution-cause-specific hazards model. We show that the cumulative incidence function can provide useful prognostic information for a particular patient but is not advisable for the primary efficacy analysis. Instead, it is preferable to fit a Cox model for the primary event which treats the competing event as an independent censoring. This is reasonably robust for controlling type I error and treatment effect bias with respect to the true primary and competing events' cause-specific hazards model, even when there is a shared, moderately prognostic, unobserved baseline frailty for the primary and competing events in that model. However, when it is plausible that a strongly prognostic frailty exists, combining the primary and competing events into a composite event should be considered. Finally, when there is an a priori interest in having both the primary and competing events in the primary analysis, we compare a bivariate approach for establishing overall treatment efficacy to the composite event approach. The ideas are illustrated by analyzing the Women's Health Initiative clinical trials sponsored by the National Heart, Lung, and Blood Institute.

Keywords

censoring; clinical benefit; cumulative incidence; hazard

1 | INTRODUCTION

Consider a 2-armed randomized clinical trial of an experimental vs control treatment where interest is focused on the treatment effect of the time to primary event that is subject to at least one competing risk. A competing risk is an event that precludes the later occurrence of the primary event. The classic competing risk is death, perhaps due to a particular cause, which precludes the later occurrence of any other event. Examples which we will discuss in

Correspondence: James F. Troendle, Office of Biostatistics Research, Division of Cardiovascular Sciences of the National Heart, Lung, and Blood Institute, NIH/DHHS, Bld RLC2 Room 9196, Bethesda, MD 20892, U.S.A. jt3t@nih.gov.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

greater detail in Section 6 are the Women's Health Initiative (WHI) estrogen + progesterone (E + P) and estrogen only (E) randomized controlled trials which randomized postmenopausal women to hormone replacement therapy or placebo.¹ The primary event for both trials was the time to coronary heart disease (CHD) which included nonfatal myocardial infarction (MI) and coronary death. A noncoronary death was therefore a competing risk for the primary event. Clearly, a noncoronary death precludes the later occurrence of a nonfatal MI or coronary death. In this paper, we will assume that there is a single competing risk which may be the first occurrence of several possible competing events.

The primary and competing events are often referred to as cause-specific events. The 2 most popular tools for analyzing the treatment effect on a cause-specific event in the presence of a competing risk are the cumulative incidence function and the cause-specific hazard (CSH) function. The cumulative incidence function is the probability that a cause-specific event occurs by a particular time. To establish some notation, let T denote the observed event time, D denote the type of event with $D = 1$ corresponding to the primary event and $D = 2$ corresponding to the competing event, and $\mathbf{X}$ be a vector of covariates. Then given $\mathbf{X}$, the cumulative incidence function for event type $D = d$ occurring by time t is $CI_d(t | \mathbf{X}) = \Pr\{T \leq t, D = d | \mathbf{X}\}$ for $d = 1, 2$.

As is well understood in the competing risks literature, the primary and competing events' cumulative incidence probabilities can impact each other in opposite directions.^{2,3} In fact, treatment can have no effect on the primary event at, say, 1 year, but the experimental-control difference in 1-year primary event cumulative incidence probabilities can suggest experimental treatment harm or benefit solely due to the treatment effect on the competing event. For this reason, we do not recommend that the primary efficacy analysis use the primary event's cumulative incidence function. Despite these reasons, some authors still advocate using group differences in cumulative incidence to evaluate treatment effects.⁴⁻⁶

As described in Koller et al.⁷ and Austin et al.,⁸ the cumulative incidence function is well suited for estimating the absolute risks of the primary and competing events which can inform a particular patient's treatment choice. For the purpose of modeling a patient's covariates' ($\mathbf{X}$) impact on the cumulative incidence function, Fine and Gray⁹ studied its hazard function called the subdistribution hazard (SDH). The SDH for cause-specific event type $D = d$ is given by

$$\begin{aligned}\lambda_d^{sd}(t | \mathbf{X}) &= -\frac{d}{dt}[\log\{CI_d(t | \mathbf{X})\}] \\ &= \lim_{\Delta t \rightarrow 0} \frac{\Pr\{t \leq T < t + \Delta t, D = d | (T > t) \cup (T < t, D \neq d), \mathbf{X}\}}{\Delta t}.\end{aligned}\quad (1)$$

As Fine and Gray⁹ note and can be seen in (1), the risk set associated with the SDH is unnatural. This is because the SDH risk set at time t for the primary event includes subjects who have previously had the competing event and so are unavailable at time t to have the primary event.

In contrast to the SDH function, the CSH function is the instantaneous probability that the cause-specific event occurs given that neither event has occurred yet. That is, the CSH function for event $D = d$ given covariates ($\mathbf{X}$) is

$$\lambda_d^{CS}(t | \mathbf{X}) = \lim_{\Delta t \rightarrow 0} \frac{\Pr\{t \leq T < t + \Delta t, D = d | T > t, \mathbf{X}\}}{\Delta t}. \quad (2)$$

Unlike the SDH, the CSH has the natural risk set which includes all subjects who have not yet experienced either the primary or competing events. Prentice et al¹⁰ showed that the likelihood under an independent censoring mechanism is completely specified by the cause-specific hazards. Thus, we believe for testing and estimating treatment efficacy, the CSH should be used instead of the SDH. This assertion will be supported by simulation studies presented in Section 3.

Although the CSH is the most relevant quantity for studying the primary event's treatment effect, it has shortcomings. The standard CSH analysis for the primary event fits a Cox model with an experimental treatment indicator term and perhaps other observed baseline covariates. In this analysis, the competing event is considered as an independent censoring when it occurs before the primary event. This analysis method is problematic when there is an unobserved, strongly prognostic subject-specific frailty that is present in both the primary and competing event's CSHs. Such a model is possible in the simulation studies presented in Section 3, where the CSH's potential difficulties are explored.

Other analysis methods should be considered when there is concern about a strong frailty in the primary and competing events' CSHs. One popular alternative is a composite event (CE) analysis which combines the primary and competing events. By definition, a CE analysis eliminates any issues with considering the competing event as a censoring. However, a major drawback to using the CE is that if treatment has little or no effect on the competing event, statistical power may be substantially lower than the power for the primary event's CSH treatment effect.

Through a careful simulation study, Freidlin and Korn¹¹ compared the CSH, SDH, and CE methods for testing the equality of the experimental vs control treatment *latent* distributions for the time to the primary event. They assumed a bivariate exponential data generating model for the joint distribution of the primary and competing events' latent times. They found that CSH-based methods were the most robust for preserving type I error rate while having good power properties. However, the latent distributions are difficult to interpret when, as in the case of the current paper, the occurrence of the competing event precludes the primary event and vice versa. We extend Freidlin and Korn's¹¹ work by generating data directly from the CSHs (Beyersmann et al¹²) without directly specifying the joint distribution of the latent times. Unlike Freidlin and Korn¹¹ who focused on testing, we also study estimation which gives insight into our type I and II error findings. We also explain where our type I and II error findings agree with theirs.

The remainder of this paper will proceed as follows. In Section 2, we describe the simulated data we will use to study the CSH, SDH, and CE approaches for estimation and testing.

Section 3 studies the CSH and SDH approaches while Section 4 examines the CE approach. In Section 5, we compare the CE approach to a simple bivariate method for testing for an overall treatment effect by separately estimating the primary and competing events' CSH or SDH ratios. Section 6 provides an illustration of the CSH and CE approaches using the WHI randomized controlled trials. Finally, in Section 7, we conclude with discussion and give our recommendations.

2 | NOTATION AND PROBLEM

Consider a 2-armed randomized clinical trial comparing an experimental vs control treatment. Let W denote the experimental treatment indicator with $W=0$ for subjects receiving the control and $W=1$ for subjects receiving the experimental treatment. For our simulations, we hypothesize that there is only one additional relevant covariate Z which is an unobserved subject-specific frailty. Therefore, the true simulation models will depend on $\mathbf{X} = (W, Z)$. The i^{th} subject is followed until 1 of 3 potential events occurs:

1. subject i has a primary event at time t_i ,
2. subject i has a competing event at time t_i and
3. subject i is lost to follow-up or the trial ends at time t_i .

We let T denote the time at which the primary or competing event occurs if either occurs before loss to follow-up, and we let D be the event type indicator with $D=1$ if the primary event occurs and $D=2$ if the competing event occurs.

The first question we consider is what is an appropriate primary analysis to determine the experimental treatment effect where interest is focused on the primary event? Next, among the appropriate approaches, what can be said about choosing the best approach for a given trial? We will investigate several primary analysis methods designed for answering these questions. The methods are broadly grouped into 2 categories: separate cause or combined cause. For each, we will briefly describe the method and its associated advantages and disadvantages. The discussion will be supported by simulation results.

Because we want to compare methods based on the CSH and the SDH functions, we simulated data in 2 different ways using the methods of Beyersmann et al.¹² The first method specifies an experimental vs control proportional CSH model for both the primary and competing events while no assumption is made on the proportionality of the SDH model for either event. This is because as shown in Beyersmann and Schumacher,¹³ the CSH and SDH are related through the equation (suppressing the dependence on the covariates $\mathbf{X}$)

$$\lambda_d^{cs}(t) = \left(1 + \frac{\Pr(T \leq t, D \neq d)}{\Pr(T > t)}\right) \cdot \lambda_d^{sd}(t). \quad (3)$$

Consequently, an experimental vs control treatment proportional CSH assumption does not imply a proportional SDH and vice versa. We call this method of simulating data the cause-specific model (CSM) method. The second method specifies a proportional SDH and proportional CSH model for the primary event, while the hazards for the competing event

may not satisfy proportional CSH or SDH. We call this method of simulating data the joint subdistribution–cause-specific model (JSD-CSM) method.

For the CSM method, we simulated data from a competing risk model assuming a log(CSH) for each of the 3 outcomes (1)–(3) given by a time-independent linear function of the experimental treatment indicator W and the frailty variable Z . The CSHs $\lambda_1^{cs}(t)$ and $\lambda_2^{cs}(t)$ for the primary event and competing event are given by (suppressing their dependence on $\mathbf{X}$)

$$\lambda_1^{cs}(t) = \exp[\alpha_1 + \beta_1 * W + \gamma_1 * Z], \quad (4)$$

$$\lambda_2^{cs}(t) = \exp[\alpha_2 + \beta_2 * W + \gamma_2 * Z], \quad (5)$$

where $Z \sim U(-\frac{\psi}{2}, \frac{\psi}{2})$, W is set to 0 for half of the subjects and 1 for the other half, and α_1 , α_2 , β_1 , β_2 , γ_1 , and γ_2 are real valued parameters ($U(a, b)$ refers to the uniform distribution on the interval $[a, b]$). Thus, the frailty Z has no effect on the primary and competing events' CSHs if and only if $\psi = 0$. We assume independent censoring with hazard given by

$$\lambda_{cen}(t) = \exp[\alpha_3]. \quad (6)$$

Beyersmann et al's¹² CSM method is quite intuitive. We generate a cause-specific event time V from the all-cause hazard $\lambda_1^{cs}(v) + \lambda_2^{cs}(v) + \lambda_{cen}(v)$ because none of the events occur prior to V . To determine which type of event occurred at time V , we perform a multinomial experiment with probabilities determined by the hazards of each event type:

$$\text{Pr(primary event)} = \frac{\lambda_1^{cs}(V)}{\lambda_1^{cs}(V) + \lambda_2^{cs}(V) + \lambda_{cen}(V)}$$

$$\text{Pr(competing event)} = \frac{\lambda_2^{cs}(V)}{\lambda_1^{cs}(V) + \lambda_2^{cs}(V) + \lambda_{cen}(V)}$$

$$\text{Pr(censoring)} = \frac{\lambda_{cen}(V)}{\lambda_1^{cs}(V) + \lambda_2^{cs}(V) + \lambda_{cen}(V)}.$$

In our simulations, the administrative censoring time is 10 years so any $V > 10$ is truncated to be an administrative censoring at 10 years. The JSD-CSM simulation method is discussed in Section D.1 of the Supporting Information.

3 | SEPARATE CAUSE METHODS FOR ASSESSING TREATMENT EFFECT

3.1 | Cumulative incidence

One popular way to assess the treatment effect on the primary event is through the experimental and control groups' cumulative incidence functions $CI_1(t|W=1)$ and $CI_1(t|W=0)$. A graphical assessment can be done by plotting the 2 cumulative incidence functions. However, as is well understood in the competing risks literature, the difference between the primary event's cumulative incidence rates $CI_1(t|W=1) - CI_1(t|W=0)$ can be substantially affected by the treatment effect on the competing event.^{2,3} In fact, even when there is no shared frailty Z in (4) and (5), which corresponds to $\psi = 0$, and there is no treatment effect on the primary event ($\beta_1 = 0$), the difference $CI_1(t|W=1) - CI_1(t|W=0)$ can suggest experimental treatment harm or benefit based solely on the treatment effect β_2 for the competing event. This is why many authors have recommended that the cumulative incidence probabilities for the primary and competing events be simultaneously assessed for patient prognosis for either of the 2 events.¹⁴ We illustrate this point through a small simulation study in Section A of the Supporting Information.

Thus, we believe that the CSH, and not cumulative incidence or the SDH, is the correct tool to assess the primary event treatment effect. A critic of such a conclusion might point out that the data for our simulation study were generated using the CSM method. However, the CSH provides the appropriate risk set for assessing the treatment effect on the primary event, ie, those patients who are at risk to have a primary event. The SDH's risk set includes such patients as well as patients who have previously experienced the competing event and so have zero probability of experiencing the primary event.

3.2 | Cause-specific hazards

To obtain a more complete assessment of the treatment effect, the CSH for the competing event must be investigated in addition to the CSH for the primary event. If one could obtain unbiased estimates of these 2 quantities, β_1 and β_2 in (4) and (5), a decision on the overall treatment effect could reasonably be made. The problem is the unavailability of such unbiased estimates when there is an unobserved shared frailty Z in (4) and (5).

To explore this phenomenon in greater detail, we used the CSM method to simulate 10 000 clinical trials each with 1000 subjects for various scenarios. The cause-specific log hazard ratio estimates $\hat{\beta}_1^{cs}$ and $\hat{\beta}_2^{cs}$ were obtained from Cox regressions using the R function `coxph` from the survival package¹⁵ where the other event was considered as a censoring, and a single term for treatment is included in the Cox model. Recall that the frailty Z is unobserved so is not available to the data analyst. Thus, the primary and competing events' cause-specific hazard functions for the analysis Cox models are

$$h_1^{cs}(t|W) = h_{01}^{cs}(t)\exp[\beta_1^{cs} * W], \quad (7)$$

$$h_2^{cs}(t|W) = h_{02}^{cs}(t)\exp[\beta_2^{cs} * W]. \quad (8)$$

We also calculated the empirical rejection rates RR_1^{cs} and RR_2^{cs} based on the Cox model Wald statistics for the respective null hypotheses

$$H_{01}^{cs}: \beta_1^{cs} = 0 \quad \text{and} \quad H_{02}^{cs}: \beta_2^{cs} = 0. \quad (9)$$

Unless $\psi = 0$ so the frailty variable $Z = 0$, the models (7) and (8) are mis-specified so $\hat{\beta}_1^{cs}$ and $\hat{\beta}_2^{cs}$ are the least false log CSHs. We also calculated subdistribution log-hazard ratio estimates $\hat{\beta}_1^{sd}$ and $\hat{\beta}_2^{sd}$ for the primary and competing events. These estimates were obtained from competing risk regressions using the R function `crr` from the `cmprsk` package¹⁵ with a single term for treatment included in the model. Thus, the analysis SDH models for the primary and competing events are

$$h_1^{sd}(t|W) = h_{01}^{sd}(t) \exp[\beta_1^{sd} * W], \quad (10)$$

$$h_2^{sd}(t|W) = h_{02}^{sd}(t) \exp[\beta_2^{sd} * W]. \quad (11)$$

We also calculated the empirical rejection rates RR_1^{sd} and RR_2^{sd} based on the model Wald statistics for the respective null hypotheses

$$H_{01}^{sd}: \beta_1^{sd} = 0 \quad \text{and} \quad H_{02}^{sd}: \beta_2^{sd} = 0. \quad (12)$$

Under the CSM, we would generally not have proportional SDHs so (10) and (11) are mis-specified and $\hat{\beta}_1^{sd}$ and $\hat{\beta}_2^{sd}$ are the least false log SDHs. The consistency of the partial likelihood estimators $\hat{\beta}_1^{cs}$, $\hat{\beta}_2^{cs}$, $\hat{\beta}_1^{sd}$, and $\hat{\beta}_2^{sd}$ is established using theorem 2.1 of Struthers and Kalbfleisch.¹⁶

The full results and discussion of the simulations using the CSM method are included in Section C of the Supporting Information and are summarized here. When the true log CSH ratios are 0 for both the primary and competing events, there is virtually no bias or type I error inflation for either the CSH or the SDH ratios. This is expected by Claim 1 in the Supporting Information which shows that the log CSH and SDH ratio estimators are all consistent for estimating 0. In contrast, when the primary event's log CSH ratio β_1 is 0 while the competing event's log CSH ratio β_2 is in the direction of experimental treatment harm, substantial bias for the estimator and type I error inflation for the test can result when using either a CSH or SDH analysis. For a CSH analysis to have type I error inflation requires a rather large shared frailty effect on the primary and competing events, while an SDH analysis in this case can have severely inflated type I error level without any shared frailty at all. In the case where experimental treatment is beneficial for both the primary and competing events, the bias increases as the frailty increases while the power is nonmonotone. In the case where experimental treatment is beneficial for the primary event

but harmful for the competing event, the primary event's CSH power increases as the frailty increases with noticeable bias for the estimator apparent when the frailty effect is large.

In Section D.3 of the Supporting Information, we provide a small simulation study in which the primary event's CSH and SDH functions in (4) and (18), respectively, have the same experimental vs control hazard ratio β_1 . There, we see similar performance for $\hat{\beta}_1^{cs}$ and $\hat{\beta}_2^{sd}$.

In Section F of the Supporting Information, we compare our results based on (4) and (5) to Freidlin and Korn's¹¹ results which are based on a bivariate latent failures model.

4 | USING THE COMPOSITE EVENT FOR THE PRIMARY EFFICACY ANALYSIS

Suppose there is pretrial clinical evidence to believe that there is an unobserved prognostic shared frailty for the primary and competing events. This could provide rationale for switching the primary efficacy analysis from the primary event to the composite event which combines the primary and competing events. The analysis Cox model for the hazard of the composite endpoint is

$$h_{comp}(t|W) = h_{0,comp}(t)\exp[\beta_{comp} * W]. \quad (13)$$

The null hypothesis is $H_{0,comp}: \beta_{comp} = 0$ which is tested using the Wald test based on the estimated log hazard ratio for the composite event $\hat{\beta}_{comp}$ obtained by fitting the Cox model (13). When there is nothing which precludes observing the composite endpoint besides independent random censoring, there will be no type I error inflation for $H_{0,comp}$.

Suppose there is no unobserved shared frailty for the primary and competing events and treatment has no effect on the competing event. In this case, the primary event log hazard ratio β_1 will be diluted in the estimated composite event log hazard ratio $\hat{\beta}_{comp}$ with a corresponding loss in power. This dilution can be quite dramatic and increases as the competing event rate increases. Table 1 shows the results of 10 000 simulated trials using the CSM method with 1000 subjects each and where $\psi = 0$ so the primary and competing events are conditionally independent given treatment.

In the top part of Table 1, $\beta_2 = 0$, so treatment has no effect on the competing event. Because $\psi = 0$, the CSM (7) is properly specified, and the competing event is correctly assumed to be an independent censoring, so $\hat{\beta}_1^{cs}$ is unbiased for $\beta_1 = -0.4$. We see that as α_2 becomes less negative, the competing event rate increases which results in fewer primary events. This causes a decrease in the power, RR_1^{cs} , for the cause-specific null hypothesis $H_{01}^{cs}: \beta_1^{cs} = 0$.

However, the decrease in power, RR_{comp} , is much more substantial for the composite event null hypothesis $H_{0,comp}: \beta_{comp} = 0$. In contrast, when as in the bottom part of Table 1, the experimental treatment is similarly efficacious for the competing event as it is for the primary event, the composite endpoint's power improves and exceeds the primary event's power. This is because competing events are counted as censoring for the primary event in a CSH analysis but as events for the composite endpoint.

5 | TESTING OVERALL TREATMENT EFFICACY

In this section, we consider the situation in which we want to include both the primary and competing events in the primary efficacy analysis. One option which we have examined is to use the composite event which includes both the primary and competing events. Another option is to use a bivariate testing rule which separately tests the primary and competing events for experimental treatment efficacy. One such rule declares the experimental treatment to be efficacious if it is superior for the primary event and noninferior for the competing event with respect to a prespecified noninferiority margin $\delta > 0$. That is, the 1-sided 97.5% upper confidence bound for the primary event's log hazard ratio is < 0 , and the 1-sided 97.5% upper confidence bound for the competing event's log hazard ratio is $< \delta$.

When we use the CSH analysis models (7)–(8), the bivariate null hypothesis we need to reject to declare overall experimental treatment benefit is

$$H_{0,biv}^{cs}: \beta_1^{cs} \geq 0 \text{ or } \beta_2^{cs} \geq \delta. \quad (14)$$

When we use the SDH analysis models (10) and (11), the bivariate null hypothesis we need to reject to declare overall experimental treatment benefit is

$$H_{0,biv}^{sd}: \beta_1^{sd} \geq 0 \text{ or } \beta_2^{sd} \geq \delta. \quad (15)$$

Analogous rules may be used to declare overall experimental treatment harm. We refer to $H_{0,biv}^{cs}$ and $H_{0,biv}^{sd}$ as win-noninferiority (WNI) null hypotheses. We reject $H_{0,biv}^{cs}$ if we reject each of its components using $\hat{\beta}_1^{cs}$ and $\hat{\beta}_2^{cs}$, similarly for $H_{0,biv}^{sd}$. We refer to these rejection rules as the CSWNI and SDWNI rules, respectively.

To study the operating characteristics of the CSWNI and SDWNI rules, we simulated data using both the CSM and JSD-CSM methods. Here, we describe only the CSM results; Section D.4 in the Supporting Information provides the JSD-CSM results. The noninferiority margin was set to $\delta = 0.3$. We simulated 10 000 clinical trials each with 1000 subjects using (4) to (6). Table 2 presents the CSWNI and SDWNI rejection rates for the WNI null hypotheses. For the purpose of comparison, we also include the composite endpoint (CE) rejection rate for the 1-sided, univariate null hypothesis $H_{0,comp}^+: \beta_{comp} \geq 0$ using model (13). Of course, the univariate null hypothesis $H_{0,comp}$ addresses a different research question from the bivariate null hypotheses $H_{0,biv}^{cs}$ and $H_{0,biv}^{sd}$.

The first set of results in Table 2 shows that when treatment has no effect on the primary or competing events, ie, $\beta_1 = \beta_2 = 0$, the type I error is appropriately controlled at 2.5% for the CE as we would expect because the composite log hazard ratio $\beta_{comp} = 0$. The type I error for the CSWNI rule is less than 2.5% because the requirement for rejection includes the additional requirement that noninferiority be met on the competing event. The SDWNI rule has an extremely small type I error rate due to negative correlation of the estimated SDH log

hazard ratios $\hat{\beta}_1^{sd}$ and $\hat{\beta}_2^{sd}$ for the primary and competing events (correlation -0.57 in the first line of Table 2).

The second set of results in Table 2 shows the power of the methods when experimental treatment is beneficial for both events. Again, the CSWNI is substantially more powerful than the SDWNI due to the negative correlation of the SDH estimates. Not surprisingly, the CE method has the highest power because experimental treatment is beneficial for both the primary and competing event components of the composite event.

The third and fourth sets of results in Table 2 show the power of the methods when experimental treatment is beneficial for the primary event but has no effect on the competing event. Again, as expected, the CSWNI is consistently more powerful than the SDWNI due to the negative correlation of the SDH estimates. As previously seen in Table 1, the CE method's power decreases with the increasing incidence of the competing event which is unaffected by treatment. The third set of results uses the same baseline hazard for the primary and competing events ($\alpha_1 = \alpha_2 = -2$) while the fourth set uses a substantially higher baseline hazard for the competing event ($\alpha_1 = -2$ and $\alpha_2 = 0$). Additional simulations (not shown) show that the CSWNI's power is highly dependent on having sufficient power for the noninferiority test.

In Section D.4 of the Supporting Information, we find in a small simulation study that the CSWNI outperforms the SDWNI when simulating data using the JSD-CSM method.

On the basis of Table 2's results and those in Section D.4 of the Supporting Information, we conclude that the CSWNI is more powerful than the SDWNI. No general conclusion can be drawn with regards to the comparison of power for the CE and CSWNI. As mentioned above, the CE addresses the univariate null hypothesis of no treatment effect on the composite endpoint while the CSWNI addresses the bivariate WNI null hypothesis (14) so the research questions are different. The research question should chiefly determine which analysis method to use.

6 | WHI CLINICAL TRIALS

The WHI remains one of the largest clinical investigations of strategies for prevention and control of aspects of morbidity and mortality in healthy, postmenopausal women having enrolled more than 161 000 women at 40 clinical centers in 1 of 3 clinical trials and an observational study.¹⁷ Responding to decades of inconclusive evidence for hormone use, between 1993 and 1998 the WHI E + P randomized controlled trial tested 16 608 women with an intact uterus for the effect of conjugated equine estrogen (0.625 mg/d) plus medroxyprogesterone acetate (2.5 mg/d) against placebo. A separate trial, the WHI E trial, tested 10 739 women without an intact uterus for the effect of conjugated equine estrogen (0.625 mg/d) against placebo. Time to CHD, defined as nonfatal MI or coronary death, was the primary outcome for both trials because hormone therapy was hypothesized to be beneficial for this outcome.

In addition to the primary CHD outcome, invasive breast cancer was designated as the primary adverse outcome. There was also a prespecified global risk index which combined the hypothesized risks and benefits of hormone therapy which included CHD, invasive breast cancer, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. Given the hypothesized risks, it can be asked whether the primary efficacy analysis should have included the competing risk of noncoronary death. The decision not to include noncoronary death was partly based on the lack of hypothesized benefit of hormone therapy for noncoronary death as well as little available evidence that noncoronary death was strongly related to CHD.

Table 3 shows the WHI E + P trial's hormone vs placebo estimated hazard ratios and associated *P* values for the cause-specific endpoints CHD and noncoronary death as well as their composite. These hazard ratios are from Cox models which included adjustment for baseline age, strata, and history of prior cardiovascular disease as was used in the primary analysis. Quite unexpectedly, hormone therapy showed significant harm for the CHD primary outcome. However, hormone therapy had no effect on noncoronary death. Consequently, had the primary outcome been the composite of CHD and noncoronary death, the trial investigators would have been unable to declare significant harm for that outcome. The dilution of the hazard ratio from 1.28 for the primary CHD outcome to 1.10 for the composite CHD + noncoronary death outcome seems related to the dilution demonstrated in Table 1.

A similar dilution is seen for the WHI E trial in Table 4. Unlike the E + P trial, the hazard ratio for the primary outcome is in the expected direction of benefit but did not reach significance.

To see how one might apply the CSWNI testing rule to the WHI data, we use the lower 95% confidence bounds for the estimated hazard ratios. To conclude harm using the CSWNI rule, the lower 95% confidence bound would have to be >1 for the primary outcome (CHD) hazard ratio and $> \exp(-\delta)$ for the competing event (noncoronary death) hazard ratio for some prespecified $\delta > 0$. Thus, for the WHI E + P trial, harm would only be concluded if the margin (on the hazard ratio scale) was at least $1 - \exp(-\delta) = 0.19$. One would probably not prespecify a margin on the hazard ratio as large as 0.19 for something as serious as noncoronary death, so the CSWNI test would fail to conclude either treatment harm or benefit. Note that using the composite event CHD + noncoronary death also would result in failure to conclude either treatment harm or benefit as the 95% CI for the composite hazard ratio includes 1. For the WHI E trial, none of the testing methods would find any significant treatment effect in the WHI E trial as the 95% CIs include 1 for both CHD and for the composite (Table 4).

7 | DISCUSSION

We have investigated different primary efficacy analyses for a 2-armed randomized clinical trial in which interest is focused on a time to primary event which may not be observed due to a competing event or independent right censoring. For our investigation, we used the method of Beyersmann et al¹² to simulate competing risks data in 2 ways. First, we

simulated from the CSHs for the primary and competing events which we called the CSM method. The CSHs (4) and (5) assumed there was an unobserved shared frailty $Z \sim U\left(-\frac{\psi}{2}, \frac{\psi}{2}\right)$ so larger values of ψ corresponded to greater patient heterogeneity. Next, we simulated from the joint subdistribution and cause-specific hazards model which we called the JSD-CSM method.

We saw that the cumulative incidence function can provide useful prognostic information but is not recommended for assessing the treatment effect. This is because the primary event's cumulative incidence increases when the competing event's cumulative incidence decreases and vice versa. We saw this phenomenon in a simulated example in which treatment had no effect on the primary event's CSH, but the experimental treatment effect on the primary event's 1-year cumulative incidence rate changed from harmful to beneficial as the experimental treatment effect on the competing event's CSH changed from beneficial to harmful.

Next we considered the primary event's CSH ratio for the efficacy analysis. We found that a near doubling of the type I error for the primary event CSH ratio estimate required a moderate harmful (respectively, beneficial) experimental treatment effect on the competing event and an average absolute frailty effect that is nearly twice the experimental treatment effect. The type I error inflation is because sicker subjects are more (respectively, less) likely to be censored in the experimental treatment group than in the control group. These findings are similar to Freidlin and Korn's¹¹ study which simulated the primary and competing event's event times from a latent failure model as opposed to a CSM. They found that a correlation of 0.5 led to a near doubling of the type I error. We concluded that the primary event's CSH should not be used for the primary analysis when there are a priori medical reasons to believe that there is a strongly prognostic, unobserved shared frailty for the primary and competing events.

When there is reason to believe that there is a shared frailty for the primary and competing events, the primary analysis can focus on the composite event which combines the primary and competing event. Although changing to the composite event moves the research question beyond the primary event, it is not subject to the type I error inflation due to the shared frailty. However, an important drawback of the composite event approach is the possible dilution of the treatment effect as compared with the primary event's CSH treatment effect. This occurs if treatment has a null or opposite effect on the competing event and is particularly problematic if the baseline hazard for the competing event is substantially higher than that of the primary event.

We saw such a dilution of the treatment effect in the WHI E + P and E trials. The primary event for both of those trials was the combination of nonfatal MI or coronary death. However, noncoronary death was not included in the primary event because it was thought to be unrelated to the primary event. In the WHI E + P trial, the primary event had a significant 1.28 hazard ratio ($P = 0.04$; Table 3). However, including noncoronary death in the endpoint diluted the hazard ratio to a nonsignificant 1.10. While the WHI E trial's primary event hazard ratio was a nonsignificant 0.91, including noncoronary death in the endpoint also diluted the hazard ratio to 0.99 (Table 4).

Finally, we investigated the bivariate WNI test for overall treatment efficacy. This is defined as a statistically significantly beneficial hazard ratio for the primary event and simultaneously a noninferiority result for the competing event. We investigated the WNI for the CSH and the SDH. We found that while both the CSH- and SDH-based WNI tests controlled the type I error substantially below the target level, the SDH test was extremely conservative due to the negative correlation of the SDH log hazard ratios for the primary and competing events. This translated into the CSH being a more powerful test. We also compared the CSH-based WNI test to the composite event's univariate test and found that the composite event test was usually more powerful. However, the exception was when there was no treatment effect on the competing event, and the baseline hazard for the competing event was at least as large as for the primary event. This finding was sensitive to the power for the noninferiority test. In any case, the bivariate research question is different from the univariate composite event research question.

In conclusion, as for any trial, the primary efficacy analysis should be appropriate to the primary research question. If this question is focused on the primary event, then the CSH should be used for the analysis as long as there are medical reasons to believe that there is not a strongly prognostic, unobserved shared frailty between the primary and competing events. However, if it is believed that such a frailty might exist, then the choice should be between either the univariate composite event test or a bivariate test such as the WNI test. Rauch and Beyersmann¹⁸ provide a useful multiple testing framework when there is an a priori interest in the composite event as well as interest in its components. If the CSH is used for the primary analysis, a post hoc sensitivity analysis should still be considered.^{19–21} Other secondary analyses can include separately examining the experimental treatment and control groups' primary and competing events' cumulative cause-specific hazards and incidence functions.¹⁴

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The views expressed in this paper are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; National Institutes of Health; or the United States Department of Health and Human Services.

We thank an associate editor and 3 reviewers for substantially improving this paper.

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (<http://biowulf.nih.gov>).

REFERENCES

1. WHI Study Group. Design of the women's health initiative clinical trial and observational study. *Control Clin Trials*. 1998;19:61–109. [PubMed: 9492970]
2. Allignol A, Schumacher M, Wanner C, Dreschler C, Beyersmann J. Understanding competing risks: a simulation point of view. *BMC Med Res Methodol*. 2011;11:86. [PubMed: 21639902]

3. Rauch G, Kieser M, Ulrich S, Doherty P, Rauch B, Schneider S, Riemer T, Senges J. Competing time-to-event endpoints in cardiology trials: a simulation study to illustrate the importance of an adequate statistical analysis. *Eur J Prev Cardiol.* 2014;21:74–80. [PubMed: 22964966]
4. Kim H Cumulative incidence in competing risks data and competing risks regression analysis. *Clin Cancer Res.* 2007;13:559–565. [PubMed: 17255278]
5. Tai B-C, Wee J, Machin D. Analysis design of randomised clinical trials involving competing risks endpoints. *Trials.* 2011;12:127. [PubMed: 21595883]
6. Campigotto F, Neuberger D, Zwicker JI. Accounting for death as a competing risk in cancer-associated thrombosis studies. *Thromb Res.* 2012;129:S85–S87. [PubMed: 22682141]
7. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? *Stat Med.* 2012;31:1089–1097. [PubMed: 21953401]
8. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. *Circulation.* 2016;133:601–609. [PubMed: 26858290]
9. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc.* 1999;94:496–509.
10. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics.* 1978;34:541–554. [PubMed: 373811]
11. Freidlin B, Korn EL. Testing treatment effects in the presence of competing risks. *Stat Med.* 2005;24:1703–1712. [PubMed: 15706579]
12. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med.* 2009;28:956–971. [PubMed: 19125387]
13. Beyersmann J, Schumacher M. Letter to the editor: comment on ‘Latouche A, Boisson V, Porcher R, and Chevret S: Misspecified regression model for the subdistribution hazard of a competing risk’. *Stat Med.* 2007;26:1649–1651. [PubMed: 17054104]
14. Latouche A, Allignol A, Beyersmann J, Lapopin M, Fine JP. A competing risks analysis should report on all-cause specific hazards and cumulative incidence functions. *J Clin Epidemiol.* 2013;66:648–653. [PubMed: 23415868]
15. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
16. Struthers CA, Kalbfleisch JD. Misspecified proportional hazards models. *Biometrika.* 1986;73:363–369.
17. Rossouw JE, Anderson G, Oberman A. Foreword. *Ann Epidemiol.* 2003;13:S1–S4.
18. Rauch G, Beyersmann J. Planning and evaluating clinical trials with composite time-to-first event endpoints in a competing risk framework. *Stat Med.* 2013;32:3595–3608. [PubMed: 23553898]
19. Slud EV, Rubinstein LV. Dependent competing risks and summary survival curves. *Biometrika.* 1983;70:643–649.
20. Scharfstein DO, Robins JM. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika.* 2002;89:617–634.
21. DiRienzo AG. Nonparametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics.* 2003;59:497–504. [PubMed: 14601750]

TABLE 1

Estimated log hazard ratios and rejection rates (%) for simulated clinical trials ^a ($N=1000$, $\alpha_1 = -2.0$, $\alpha_3 = -1.5$, $\gamma_1 = \gamma_2 = 1.0$, $\beta_1 = -0.4$, $\Psi = 0$)

α_2	β_2	$\hat{\beta}_1^{csc}$	RR_1^{csc}	$\hat{\beta}_{comp}^b$	Relative dilution ^d (%)	RR_{comp}^c
-4.0	0.0	-0.402	93.6	-0.344	14	89.94
-3.0	0.0	-0.401	91.8	-0.276	31	79.63
-2.0	0.0	-0.400	86.4	-0.179	55	53.19
-1.0	0.0	-0.401	70.1	-0.093	77	22.33
-1.0	-0.2	-0.401	72.2	-0.250	38	89.4
-1.0	-0.4	-0.399	73.5	-0.400	0	99.9

^aEstimated over 10 000 replicated trials; target rejection rate is 5% under null.

^bStandard error of estimates over the 10 000 replications is .001.

^cRR is rejection rate.

$$^d \text{Relative dilution} = 100 * \left(1 - \frac{\text{average}(\hat{\beta}_{comp})}{\beta_1} \right)$$

TABLE 2

Rejection rates (%) for simulated clinical trials from cause-specific models, to (4) to (6)^a ($N = 1000, \alpha_1 = -2.0, \alpha_3 = -1.5, \gamma_1 = \gamma_2 = 1.0, \text{NI margin} = 0.3$)

Ψ	α_2	β_1	β_2	CE ^b	CSWNI ^c	SDWNI ^d
0	-2	0	0	2.36	1.79	0.48
1	-2	0	0	2.32	2.09	0.44
3	-2	0	0	2.32	2.04	0.31
0	-2	-0.4	-0.4	99.29	87.42	55.04
1	-2	-0.4	-0.4	99.81	91.56	46.79
3	-2	-0.4	-0.4	97.74	79.99	24.55
0	-2	-0.4	0	53.19	60.82	25.26
1	-2	-0.4	0	58.04	70.90	22.59
3	-2	-0.4	0	45.31	66.62	13.87
0	0	-0.4	0	8.18	42.09	32.72
1	0	-0.4	0	8.85	45.08	35.29
3	0	-0.4	0	6.88	45.56	36.97

^aEstimated over 10 000 replicated trials; target rejection rate is 2.5% under the null hypothesis.

^bCE is the 1-sided test of the composite endpoint null hypothesis $H_{0,comp}$.

^cCSWNI is the test of $H_{0,biv}^{cs}$ using cause-specific hazard ratios and the win-noninferiority rejection region.

^dSDWNI is the test of $H_{0,biv}^{sd}$ using subdistribution hazard ratios and the win-noninferiority rejection region.

TABLE 3

Estimated HR for WHI E+P clinical trial

Outcome	# Events	$\widehat{HR}$	95%CI	P value
CHD + noncoronary death	656	1.10	(0.94,1.29)	.23
CHD = primary outcome	286	1.28	(1.01,1.62)	.04
Noncoronary death	370	0.98	(0.81,1.19)	.82

Abbreviations: CHD, coronary heart disease; E + P, estrogen + progesterone; HR, hazard ratio; WHI, Women's Health Initiative

TABLE 4

Estimated HR for WHI E clinical trial

Outcome	# Events	$\widehat{HR}$	95%CI	P value
CHD + noncoronary death	817	0.99	(0.86,1.14)	.94
CHD = primary outcome	376	0.91	(0.75,1.11)	.35
Noncoronary death	441	1.07	(0.88,1.30)	.46

Abbreviations: CHD, coronary heart disease; E, estrogen only; HR, hazard ratio; WHI, Women's Health Initiative