Published in final edited form as:

Biometrics. 2018 September ; 74(3): 1006–1013. doi:10.1111/biom.12871.

A Group Sequential Test for Treatment Effect Based on the Fine-Gray Model

Michael J. Martens and

The Emmes Corporation, Rockville, Maryland, USA

Brent R. Logan

Division of Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

Summary

Competing risks endpoints arise when patients can fail therapy from several causes. Analyzing these outcomes allows one to assess the direct benefit of treatment on a primary cause of failure in a clinical trial setting. Regression models can be used in clinical trials to adjust for residual imbalances in patient characteristics, improving the power to detect treatment differences. But, none of the competing risks methods currently available for use in group sequential trials adjust for covariates. We propose a group sequential test for treatment effect that, because it is based on the Fine-Gray model, permits adjustment for covariates. Our derivations show that its sequence of test statistics has an asymptotic distribution with an independent increments structure, which allows standard techniques such as O'Brien-Fleming designs and error spending functions to be employed to meet type I error rate and power specifications. We demonstrate the test in a reanalysis of BMT CTN 0402, a phase III clinical trial that evaluated an experimental treatment for the prevention of adverse outcomes following blood and marrow transplant. Moreover, using a simulation study of randomized group sequential trials, we demonstrate that the proposed method preserves the type I error rate and power at their nominal levels in the presence of influential covariates.

Keywords

Competing Risks Analysis; Covariate Adjustment; Fine-Gray Regression Model; Graft Versus Host Disease; Group Sequential Design; Hematopoietic Cell Transplantation

1. Introduction

A late phase clinical trial is conducted with the purpose of comparing efficacy of an experimental treatment to a control. Because these trials can be extremely expensive and lengthy, means for improving the efficiency of this treatment evaluation hold great value. One method is group sequential design, which can provide substantial reductions in sample

Supplementary Materials

Web Appendices A and B, referenced in Section 2.3, are available with this paper at the *Biometrics* website on Wiley Online Library. The GVHD data from Section 4 and a SAS program implementing the group sequential Fine-Gray test on these data are also available here.

size and trial duration by permitting early stopping of the trial for efficacy and/or futility. But, choosing the stopping boundaries in a way that will provide control of the type I error rate for the trial requires knowledge of the joint distribution of the test statistics. When this distribution has an independent increments structure (Jennison and Turnbull, 1999), standard methods can be utilized to specify appropriate boundaries, including Pocock and O'Brien-Fleming tests and error spending functions (Pocock, 1977) (O'Brien and Fleming, 1979) (Lan and DeMets, 1983). Adjustment for covariates can also deliver more efficient assessment of the treatment effect. Analysis approaches that offer this benefit include regression models and stratified methods. They improve efficiency by taking into account covariates that impact the outcome of interest and might confound treatment evaluation. This is valuable even in randomized clinical trials, where sizable imbalances on influential covariates may still occur between treatment groups purely by chance (Ciolino et al., 2015).

Time to event outcomes are often the main interest of a clinical trial. For example, an oncology trial may evaluate a new treatment to determine whether it provides better overall survival for patients compared to usual care. Therefore, a strong need exists for time to event analysis methods that incorporate these two techniques, group sequential design and covariate adjustment. This is true both for survival outcomes, where patients are at risk of a single event, and for competing risks outcomes, where they might experience one of multiple events. In the survival data setting, the log-rank test and the Cox proportional hazards model (Cox, 1972) each have been shown to have an independent increments structure (Tsiatis, 1981) (Bilias et al., 1997), with the latter permitting covariate adjustment.

When competing risks are present, interest may lie in whether a treatment reduces the risk of occurrence of a specific type of event. As an example, consider the Blood and Marrow Transplant Clinical Trials Network (BMT CTN) study 0402, a phase III clinical trial that evaluated an experimental treatment for the prevention of graft versus host disease (GVHD) after blood and marrow transplantation. The development of GVHD often results in an increased risk of mortality; however, because a patient may die before GVHD develops, death before GVHD is a competing risk for the event of primary interest, onset of GVHD.

In this setting, a competing risks analysis method is appropriate. Gray's test and a pointwise comparison of cumulative incidence have the independent increments structure, as well (Logan and Zhang, 2013); however, neither of these methods adjusts for covariates. Previously, no competing risks method that adjusts for covariates had been shown to have an independent increments structure. As shown in Scharfstein et al. (1997), an analysis method that is semiparametric efficient for the parameter of interest must give a sequence of test statistics that converges to an independent increments structure in a group sequential setting. The widely used Fine-Gray regression model (Fine and Gray, 1999) can perform covariate adjustment, but is not known to provide semiparametric efficient estimation of its parameter vector. Therefore, a direct approach may be required to derive the joint distribution of its test statistics and determine if it possesses independent increments. This manuscript presents a test for treatment effect on a competing risks outcome, derived from the Fine-Gray model, that has this structure and is readily applicable in a group sequential trial.

The sequel is organized as follows. Section 2 introduces the Fine-Gray model and its parameter estimating process in the group sequential setting, presents the asymptotic distribution of test statistics for this process, and proposes a group sequential test for treatment effect. Section 3 displays results of a simulation study that investigates the small sample performance of this test with regard to meeting type I error rate and power specifications for a randomized group sequential trial. Furthermore, these simulations evaluate whether conditional power of this test accurately reflects its true rejection probability. In section 4, we demonstrate the proposed test through a redesign of BMT CTN 0402. We summarize these results and discuss extensions of our methods in Section 5.

2. Methods

2.1 Competing Risks Data Notation

We consider competing risks data in a clinical trial from a random sample of patients whose event times might be right-censored due to loss to follow-up. In the group sequential setting, both the set of patients accrued to the study and the data available on these patients can differ between analyses, so the calendar times at which patients are accrued must be considered. For the i^{th} patient, let T_i denote the event time, e_i the event type, C_i the loss to follow-up censoring time, τ_i the calendar time of accrual, and $\mathbf{Z}_i$ a vector of covariates of length p . More specifically, T_i is the time to the first competing event, $e_i \in \{1, \dots, J\}$ where J is the number of event types, τ_i is the calendar time of accrual relative to the start of enrollment, and the covariates $\mathbf{Z}_i$ are time-invariant.

In a group sequential trial, we have two time scales to consider: the calendar time t , and the survival time s . We assume an upper bound t^* exists on the duration of the study. An event can be observed for patient i at survival time s only if he has been on study for s units of time, i.e. only if $t - \tau_i \geq s$. Therefore, we only consider pairs $(t, s) \in D$, where $D = \{(t, s): 0 \leq s \leq t - \tau_i^+, t \leq t^*\}$. Let $v^+ = \max\{v, 0\}$ denote the positive part of a scalar v . The observed data for patient i at calendar time t is $(X_i(t), \Delta_i(t), \Delta_i(t)e_i, \tau_i, \mathbf{Z}_i)$, where $X_i(t) = T_i \wedge C_i \wedge (t - \tau_i)^+$ is the observed survival time and $\Delta_i(t) = I\{T_i \leq C_i \wedge (t - \tau_i)^+\}$ is the event indicator at calendar time t . $\Delta_i(t)$ and $\Delta_i(t)e_i$ are included in the observations, since a patient's event type e_i is known at calendar time t only if they are not censored. The variable $B_i(t) = C_i \wedge (t - \tau_i)^+$ represents the effective censoring time at calendar time t for patient i as the minimum of the loss to follow-up censoring time C_i and the administrative censoring time $(t - \tau_i)^+$.

2.2 Fine-Gray Regression Model with Staggered Entry

A primary quantity of interest in competing risks analysis is the cumulative incidence function of cause j , defined as

$$F_j(s) = P(T \leq s, \varepsilon = j)$$

for each event type $j = 1, \dots, J$. This is a subdistribution function representing the probability of experiencing a cause j event by time t in the presence of all other event types. It has a direct relationship to the sub distribution hazard function of cause j ,

$$\lambda_j^*(s) = -\frac{\partial}{\partial s} \log\{1 - F_j(s)\} = \lim_{h \rightarrow 0} \frac{P\{s \leq T \leq s+h, e=j | T \geq s \text{ or } (T < s, e \neq j)\}}{h}.$$

This one-to-one relationship also holds between $\lambda_j^*(s|\mathbf{Z})$ and $F_j(s|\mathbf{Z})$, the sub distribution hazard and cumulative incidence functions conditional on covariates.

Without loss of generality, we assume that just two causes of failure exist and that we are evaluating the treatment effect on cause 1 events. One method for performing this assessment while accounting for the impact of other covariates is the Fine-Gray model (Fine and Gray, 1999)

$$\lambda_1^*(s|\mathbf{Z}) = \lambda_{10}^*(s) \exp(\beta' \mathbf{Z}) \text{ for } s \leq t,$$

where Z_1 , the first component of $\mathbf{Z}$, is the treatment indicator, coded as 1 and 0 for the experimental and control treatments, respectively. Because the subdistribution hazard function has a one-to-one relationship with its corresponding cumulative incidence function, a test of $H_0: \beta_1 = 0$ will determine whether the cumulative incidence of cause 1 differs between treatment groups, where β_1 is the first component of β .

We make the following assumptions about the trial data for the observations $i = 1, \dots, n$:

1. T_i and C_i are continuous random variables.
2. The Fine-Gray model holds for cause 1 events.
3. The $(T_i, C_i, \varepsilon_i, \tau_i, \mathbf{Z}_i')$ are independently and identically distributed.
4. Independent censoring, i.e. $C_i \perp\!\!\!\perp (T_i, \varepsilon_i, \tau_i, \mathbf{Z}_i')$.
5. Independent accrual, i.e. $\tau_i \perp\!\!\!\perp (T_i, C_i, \varepsilon_i, \mathbf{Z}_i')$.

Assumptions 3 and 4 are made in the standard proportional subdistribution hazards model of Fine and Gray (1999). The fifth assumption is made by Sellke and Siegmund (1983) in their derivation of a sequential test for the Cox model.

2.3 Parameter Estimation

Similar to the Cox proportional hazards model, a score process is used to estimate β . In the group sequential setting, this process has the form

$$U(\beta, t, s) = \sum_{i=1}^n \int_0^s \{Z_i - \bar{Z}(\beta, t, u)\} w_i(t, u) dN_i^1(u),$$

where $N_i^1(s) = I(T_i \leq s, \varepsilon_i = 1)$ is the counting process of cause 1 events for patient i , $Y_i^1(s) = 1 - N_i^1(s^-)$ is the “at risk” process of cause 1 events for patient i ,

$\bar{Z}(\beta, t, s) = \sum_{i=1}^n w_i(t, s) Y_i^1(s) Z_i \exp(\beta' Z_i) / \sum_{i=1}^n w_i(t, s) Y_i^1(s) \exp(\beta' Z_i)$, and the $w_i(t, s)$ are inverse probability of censoring weights (IPCW). These weights have the form $w_i(t, s) = I\{B_i(t) \geq X_i(t) \wedge s\} \hat{G}(t, s) / \hat{G}(t, X_i(t) \wedge s)$, where $\hat{G}(t, s)$ is the Kaplan-Meier estimator of the survival function $G(t, s) = P[B_i(t) \geq s]$ of $B_i(t)$. It is important to note that in the presence of censoring, N_i^1 may not be fully observed, but $w_i N_i^1$ is. The definition of Y_i^1 matches that in Fine and Gray (1999), as the proper risk set for the subdistribution hazard includes both survivors of all events and those who experienced a competing event. When estimation is performed at the single time $s = t = t^*$, these weights coincide with those used in Fine and Gray (1999).

Let β_0 denote the true value of the regression parameter. We estimate β_0 with the value $\hat{\beta}$ that solves $U(\hat{\beta}, t, s) = 0$. This can be done for every $(t, s) \in D$, giving us an estimating process $\hat{\beta}(t, s)$ for β_0 . The second of the following theorems provides the asymptotic distribution of this process.

Theorem 1—Under regularity conditions 1–4, $n^{-1/2} U(\beta_0) \xrightarrow{d} \xi$, where ξ is a zero mean Gaussian process with continuous sample paths and covariance function

$$E\{\xi(t_1, s_1) \xi(t_2, s_2)'\} = \int_0^{s_1 \wedge s_2} G(t_1 \wedge t_2, u) \{\Gamma_2(u) - \Gamma_1(u) \otimes 2 / \Gamma_0(u)\} \lambda_{10}^*(u) du,$$

where $\Gamma_j(s) = E\{Y_i^1(s) Z_i^{\otimes j} \exp(\beta' Z_i)\}$ for $j = 0, 1, 2$.

Theorem 2—Under regularity conditions 1–5, $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \zeta$, where ζ is a zero mean Gaussian process with continuous sample paths and covariance function

$$E\{\zeta(t_1, s_1) \zeta(t_2, s_2)'\} = E\{\xi(t_1, s_1) \otimes 2\}^{-1} E\{\xi(t_1, s_1) \xi(t_2, s_2)'\} E\{\xi(t_2, s_2) \otimes 2\}^{-1}.$$

Proofs of these results and statement of the regularity conditions on which they rely can be found in Web Appendices A and B. The derivations rely heavily on empirical process theory presented in Pollard (1990) and Kosorok (2008), following a similar strategy as that used in Biliak et al. (1997).

In practice, a researcher will choose to use all available follow-up information on patients to estimate β_0 at each analysis, which is done by setting $s = t$. Then, $\hat{\beta}(t, t)$ and the asymptotically normal distribution of $\sqrt{n}\{\hat{\beta}(t, t) - \beta_0\}$ are employed to draw inference on β_0 at calendar time t .

2.4 Group Sequential Test for Treatment Effect

A group sequential test permits early stopping of a study for efficacy or futility. Because the same data are used for testing at multiple stages, the joint distribution of the test statistics across analyses must be known in order to control the overall type I error rate. For many common analysis methods, this joint distribution follows exactly or asymptotically a specific multivariate normal distribution called the canonical distribution (Jennison and Turnbull, 1999). The sequence of statistics $(W_1, \dots, W_K)$ is said to follow an independent increments structure with the information sequence $(I_1, \dots, I_K)$ for θ if its asymptotic distribution is multivariate normal with $W_j \sim \mathcal{N}(\theta, 1)$ and $\text{Cov}(W_i, W_j) = \sqrt{I_i/I_j}$ for $i < j$; this distribution is the canonical distribution.

Theorem 2 implies that $\sqrt{n}\{\hat{\beta}(t, t) - \beta_0\}$ has independent increments at calendar times $t_1, \dots, t_K$. This further implies that its first component, $\sqrt{n}\{\hat{\beta}_1(t, t) - \beta_{01}\}$, also has independent increments. Fine and Gray (1999) gives a consistent variance estimator for $\sqrt{n}\{\hat{\beta}(t, t) - \beta_0\}$; by considering its $(1, 1)$ component, we get a consistent variance estimator for $\sqrt{n}\{\hat{\beta}_1(t, t) - \beta_{01}\}$. Therefore, a group sequential Wald test of treatment effect (i.e., whether $\beta_1 = 0$) can be obtained from these results by comparing the standardized test statistics $W_j = \hat{\beta}_1(t_j, t_j) / \text{SE}\{\hat{\beta}_1(t_j, t_j)\}$, calculated at interim analysis times t_j to appropriate stopping boundaries. Because we are considering a competing risk endpoint and the information for $\hat{\beta}_1$ is approximately proportional to the number of cause 1 events (Latouche et al., 2004), the observed information levels at interim analyses are unpredictable and can deviate greatly from the levels expected at the design stage. Therefore, an error spending approach to group sequential testing is more appropriate than a Pocock or O'Brien-Fleming test, which each assume equal increments in information between analyses.

We can also derive a group sequential test for the situation where interest lies in comparing cumulative incidence on the interval $[0, s]$, where s is fixed and $s < t^*$. For instance, a comparison of disease relapse during the first 2 years following bone marrow transplant between treatment groups may be desired, using a trial of a 5 year duration. In this case, the one parameter process $\{\hat{\beta}(t, s) : t \in [s, t^*]\}$ is used for inference, with its asymptotic distribution provided by our derivation. This distribution also possesses independent increments, so a group sequential test in this setting can be obtained similar as before by using the test statistics $\hat{\beta}_1(t_j, s) / \text{SE}\{\hat{\beta}_1(t_j, s)\}$.

3. Simulation Study

3.1 Design

We performed a simulation study to assess how well the proposed test performed with regards to meeting type I error and power specifications for a group sequential trial. We simulated randomized clinical trials that compared efficacy of an experimental treatment to a control through two-sided testing of the existence of a treatment effect ($H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$). Treatment assignments were randomized at a 1:1 ratio. Each trial consisted of an

accrual period $[0, A]$ during which patients are accrued to the study at a uniform rate. The entire study duration is denoted by L , giving a minimum planned follow-up time of $L-A$ for each patient. Group sequential testing was performed at the 0.05 level using the proposed method and Gray's test (Gray, 1988) with a three stage design employing the alpha spending function $0.05 \min\{1, IF^3\}$, where IF is the information fraction, the fraction of total information for the trial. Interim analyses were prespecified at calendar times where we expect information fractions of $1/3$ and $2/3$, corresponding to $1/3$ and $2/3$ of the total number of cause 1 events.

Data was generated in a manner so that the cumulative incidence functions were

$$F_1(s|\mathbf{Z}) = 1 - \left\{1 - q(1 - e^{-0.4s})\right\}e^{\beta'\mathbf{Z}} \quad \text{and} \quad F_2(s|\mathbf{Z}) = (1 - q)e^{\beta'\mathbf{Z}} \left\{1 - \exp(-0.3se^{\gamma'\mathbf{Z}})\right\}.$$

The covariates $\mathbf{Z}$ impact both probabilities, with β indicating γ their effects on the risk of cause 1 and cause 2 failures, respectively. This implies that the Fine-Gray model in Section 2.2 holds for cause 1 events.

Choices of simulation parameters consisted of the following: sample sizes per treatment group, $n = 100, 200$, or 400 ; proportion of cause 1 events, $q = 25\%$, 50% , or 75% ; accrual and maximum study periods, $(A, L) = (2, 5)$ or $(3, 4)$; independent censoring, exponentially distributed at a rate of 5% or 10% per year; $\beta_1 = 0$ or δ , a treatment effect at which the proposed test should have 80% power under its test statistics' asymptotic distribution; and $\phi = \log 1.5$ or $\log 2$, where ϕ is a parameter describing the strength of the covariates' effects. δ was computed for each specification using the sample size formula for the Fine-Gray model proposed in (Latouche et al., 2004), which computes a total sample size using the specified type I error rate, power, allocation ratio of treatment groups, proportion of cause 1 events observed, and δ .

We considered three scenarios, each corresponding to a specification of these effects. Scenario 1 had one $N(0, 1)$ covariate considered with $\beta_2 = \phi$, $\gamma = 0$. Scenario 2 had two covariates, one $N(0, 1)$ and one $Bernoulli(0.5)$, with $\beta_2 = \beta_3 = (2/\sqrt{5})\phi$ and $\gamma = 0$. And, scenario 3 had two covariates, both $N(0, 1)$, with $\beta_2 = \gamma_3 = \phi$ and $\beta_3 = \gamma_2 = 0$. All scenarios involve a covariate that influences the failure type of interest. Scenario 2 has two such variables, while the other scenarios have one. Moreover, scenario 3 has a covariate that affects cause 2 but has no impact on cause 1 and, therefore, should not impact the proposed test. The size of covariate effect(s) for the scenarios were selected so that the linear predictor $\beta'\mathbf{Z}$ will have the same mean and variance and, thus, a similar effect on the cause 1 subdistribution hazard for all scenarios.

Simulations were performed in R software using the package "cmprsk" to compute the test statistics for the Fine-Gray model and Gray's test and the package "gsDesign" to compute the efficacy stopping boundaries for the three stages using the specified alpha spending function.

3.2 Type I error and power

10000 trials were generated for each scenario, under each combination of simulation parameters. Using the results, we obtained Monte Carlo estimates of the cumulative type I error rates and power over the three stages of the group sequential design for both the proposed test and Gray's test. We compare the estimates of stagewise cumulative type I error rates in Figure 1. To summarize the large number of simulation results (144 parameter combinations $\times$ 3 scenarios = 432 configurations), we collect these estimates across choices of q , (A, L) , censoring rate, and ϕ into box plots that describe the type I error rate for each choice of n . The nominal cumulative type I error rates under the asymptotic distributions are shown by the black lines. We see that for all three stages, the estimated rates fall close to the nominal rates for both Gray's test and the Fine-Gray test. For a given stage, the width of a box plot has an expected size if the test statistics have converged to their asymptotic distribution. The sizes of the box plots are similar to this expectation for all stages and tests except for the Fine-Gray test at stage 1 with sample sizes of 100 per group. We attribute this to the small number of cause 1 events observed on average in stage 1 for these sample sizes, which is likely insufficient for convergence of the parameter estimates to their asymptotic normal distribution.

Figure 2 illustrates the impact of the influential covariates' effects on cumulative power at the three stages. Again, each box plot is constructed from estimates obtained from configurations with the corresponding ϕ value. Nominal levels of cumulative power are shown by the black lines. At stage 1, both analysis methods fall short of the nominal power, with roughly equivalent values. On the other hand, at stages 2 and 3, the Fine-Gray test comes close to the nominal power levels, while Gray's test falls short because it fails to adjust for covariates. Also, a greater loss of power occurs for Gray's test with the larger value of ϕ , when the covariates exert more influence over the probability of cause 1 events. We generally see 5%–10% higher power for the Fine-Gray test over Gray's test in the simulation scenarios considered.

3.3 Conditional power

Conditional power, the probability of rejecting H_0 under a given value of the target parameter using available data, is sometimes used as a method for stochastic curtailment of a trial, with low conditional power used as a criterion for early stopping for futility. The parameter value at which conditional power is computed is often its value under the null hypothesis, alternative hypothesis, or an estimate computed from the data. For a continuous outcome, Ciolino et al. (2014) demonstrated that the conditional power given by a t-test, an unadjusted analysis method, is dependent on the level of imbalance of influential covariates between treatment groups and that this dependency can inflate the probability of wrongly stopping a trial early when conditional power is used for curtailment.

We performed a similar evaluation in the competing risk setting by investigating whether the conditional power was an accurate representation of the probability of ultimately rejecting H_0 . For simplicity, we consider the setting where one influential covariate is at play, given in scenario 1. Ciolino et al. (2015) showed that the t statistic of imbalance, defined as

$$t_{imb} = \frac{\bar{Z}_{2,1} - \bar{Z}_{2,0}}{s_z \sqrt{n_1^{-1} + n_0^{-1}}},$$

is a useful summary measure of the imbalance in covariate distributions between treatment groups, where $\bar{Z}_{2,i}$ is the sample mean of Z_2 in subjects with $Z_1 = i$, n_i is the number of subjects with $Z_1 = i$, and s_z is the pooled variance of both treatment groups. Conditional power to detect $\beta_1 = \theta$ at analysis k is defined as $CP_k(\theta) = P_{\theta}(\text{reject } H_0 | W_k)$. By integrating over W_k , we get the unconditional power for the test, $P_{\theta}(\text{reject } H_0) = E\{CP_k(\theta)\}$. The question is whether this conditional power accurately represents the probability of rejecting H_0 ; we suspect that covariate imbalance may still affect the probability of rejection. Therefore, using the simulation results for scenario 1, we fit the logistic regression models

$$\text{logit } P_{\theta}(\text{reject } H_0 | t_{imb,k}) = \text{logit } CP_k(\theta) + v_1 t_{imb,k} + v_2 |t_{imb,k}|$$

for stages $k = 1, 2$ and for $\theta = 0$ and δ , its values under the null and alternative hypotheses. Here, $t_{imb,k}$ denotes the t statistic of imbalance in the set of patients accrued by interim analysis k . In preliminary simulations that assessed performance of Gray's test in the presence of an influential covariate, we found that $|t_{imb}|$ impacted the type I error rate, while t_{imb} adversely affected its power. Hence, we include both as predictors in these models. We use box plots to summarize the results of these models under the various choices of simulation parameters. These are plots of the parameter estimates $\hat{v}_1$ and $\hat{v}_2$. If the conditional power does accurately describe the ultimate rejection probability, these estimates and their corresponding box plots should be close to 0. But, if the imbalance impacts the conditional likelihood of rejecting the null hypothesis, we expect $\hat{v}_1$ and $\hat{v}_2$ to differ from 0.

Coefficient estimates for $|t_{imb}|$ under the null hypothesis are shown in Figures 3 (a)–(b) for stages 1 and 2, respectively. The box plots for the Fine-Gray test are centered near 0, whereas those for Gray's test show a positive bias that increases as ϕ does. This implies that imbalance on an influential covariate between treatment groups will cause the conditional power of Gray's test at $\beta_1 = 0$ to underestimate the true type I error rate with increasing severity as the size of the imbalance grows.

Figures 3 (c)–(d) illustrate a similar trend for t_{imb} under the alternative $\beta_1 = \delta$ at stages 1 and 2. The coefficient for t_{imb} is close to 0 for the proposed test but has a negative bias for Gray's test. When a negative imbalance exists between groups with respect to their covariate distributions, the conditional power of Gray's test at $\beta_1 = \delta$ will tend to underestimate the true power of this test; moreover, this inaccuracy will increase as ϕ does. These results imply that if low conditional power is used as a criterion for curtailment, the trial will be more likely to stop wrongly for futility when a treatment effect actually exists. Our simulations permitted early stopping for efficacy only; however, this finding also illustrates the detrimental effect that covariate imbalance could have if futility stopping were included.

Namely, a negative imbalance would inflate the stagewise type II error rates and produce an underpowered test.

4. Example

To demonstrate the use of the proposed test, consider the Blood and Marrow Transplant Clinical Trials Network 0402 clinical trial (Cutler et al., 2012) that evaluated the efficacy of a proposed prophylactic treatment for graft versus host disease (GVHD). The experimental treatment, a combined regimen of Tacrolimus and Sirolimus (Tac/Sir), was compared to the current standard regimen of Tacrolimus and Methotrexate (Tac/Mtx) in a randomized clinical trial of 304 patients. The treatment groups did not differ significantly on the primary endpoint, acute GVHD-free survival, a composite endpoint indicating that a patient neither suffered acute GVHD nor died. Though the trial did not demonstrate superiority of the experimental treatment, patients in the Tac/Sir arm fared no worse than those in the Tac/Mtx arm on this and several secondary endpoints, suggesting that Tac/Sir is a satisfactory alternative to Tac/Mtx.

We will reanalyze the trial data to assess the treatment effect on extensive chronic GVHD (eCGVHD), which was not assessed in the original trial. This is a competing risks outcome that has death before eCGVHD as a competing event. The variables Karnofsky performance score (≥ 90 vs. < 90), donor-recipient gender matching (female donor/male recipient vs. others), and recipient age (0–40 vs. 40–50 vs. > 50 years) are suspected to influence the risk of GVHD. Therefore, we will use the proposed test to account for the potential influence of these covariates on eCGVHD and compare the results to those from Gray's test, which makes no such adjustment.

We use group sequential Gray's and Fine-Gray tests for treatment evaluation in a three stage design with interim analyses performed at calendar times when $1/3$ and $2/3$ of the total number of eCGVHD events, 139, have been observed. The error spending functions from the power family in Jennison and Turnbull (1999) with $\rho = 1$ or 3 are employed for efficacy stopping, which give stopping boundaries similar to the Pocock and O'Brien-Fleming designs. Gray's test gives the standardized score statistics $W_{g1} = 2.213$, $W_{g2} = 2.310$, and $W_{g3} = 2.090$; and the Fine-Gray tests gives standardized statistics $W_{f1} = 2.134$, $W_{f2} = 2.209$, and $W_{f3} = 1.961$ derived from $\hat{\beta}_1$ and its variance estimator at each stage.

Figure 4 displays the stopping boundaries and test statistics for the alpha spending functions and tests considered; only the upper boundaries are shown, as the lower boundaries are symmetric to them about the x-axis. The $\rho = 1$ error spending function gives stagewise critical values of $c_1 = 2.394$, $c_2 = 2.294$, and $c_3 = 2.200$; and the $\rho = 3$ function gives values of $d_1 = 3.113$, $d_2 = 2.462$, and $d_3 = 2.009$. If the $\rho = 1$ function is used, Gray's test will reject the null hypothesis at stage 2, while the Fine-Gray test will accept the null. Under the $\rho = 3$ family, Gray's test will reject at stage 3, while the Fine-Gray test will again accept the null hypothesis of no treatment effect. With either spending function, Gray's test would assert that Tac/Sir was inferior to Tac/Mtx in the prevention of eCGVHD, while the Fine-Gray test would find no compelling evidence of a difference in the efficacy of these treatments.

5. Discussion

We introduced a group sequential test for treatment effect on a competing risk endpoint that makes adjustment for influential covariates. A comprehensive simulation study demonstrated that this test properly adjusts for these covariates by maintaining type I error and power at the nominal levels. Because the test statistics produced by this method have an independent increments structure, standard methods can be used to specify stopping boundaries. Furthermore, this is the first group sequential analysis method that utilizes IPCW.

The derivations and simulations for this test were performed under the standard assumptions of the Fine-Gray model; that is, we assumed proportionality of the subdistribution hazard over the study duration and that the censoring distribution is independent of covariates. However, situations do exist where these assumptions are violated. For example, patients in poorer health could be more likely to drop out of a study. He et al. (2016) demonstrated that the Fine-Gray model can produce biased estimates when covariate-dependent censoring is present. The bias is introduced by the estimates of the censoring distribution that are used in the IPCW weights of the score process. To address this problem, they utilized a Cox proportional hazards model to estimate the censoring distribution conditional on these covariates. Simulations demonstrated that this method can reduce the bias even when the censoring distribution does not truly follow a Cox model. We may be able to incorporate this adjustment of the censoring estimates into our group sequential Fine-Gray test.

In the case of nonproportional subdistribution hazards, one alternative approach is to assess cumulative incidence of the failure type of interest at a fixed point in time. The direct binomial regression technique of Scheike et al. (2008) and pseudo-value regression of Klein and Andersen (2005) are two methods capable of this. Neither has been applied to the group sequential setting, though. In particular, direct binomial regression bears many similarities to the Fine-Gray model, and a similar group sequential test of treatment effect may be derived from this method to comparing cumulative incidence at a fixed time point while adjusting for other factors. We are developing such a group sequential test, which will be presented in a future report.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support for this study was provided by grant #F31HL134317 to Michael J. Martens by the National Heart, Lung, and Blood Institute of the National Institutes of Health. Support for the BMT CTN 0402 trial was provided by grant #U10HL069294 to the Blood and Marrow Transplant Clinical Trials Network from the National Heart, Lung, and Blood Institute and the National Cancer Institute, along with contributions by Wyeth Pharmaceuticals Inc. The authors thank the BMT CTN for permitting the use of the 0402 data for the example. The content is solely the responsibility of the authors and does not necessarily represent the official views of the above mentioned parties.

References

- Bilias Y, Gu M, Ying Z. Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics*. 1997; 25:662–682.
- Ciolino JD, Martin R, Zhao W, Jauch EC, Hill MD, Palesch YY. Continuous covariate imbalance and conditional power for clinical trial interim analyses. *Contemporary clinical trials*. 2014; 38:9–18. [PubMed: 24607294]
- Ciolino JD, Martin RH, Zhao W, Hill MD, Jauch EC, Palesch YY. Measuring continuous baseline covariate imbalances in clinical trial data. *Statistical methods in medical research*. 2015; 24:255–272. [PubMed: 21865270]
- Cox D. Regression and life tables (with discussion). *Journal of the Royal Statistical Society Series B*. 1972; 34:187–220.
- Cutler C, Logan BR, Nakamura R, Johnston L, Choi SW, Porter DL, Hogan WJ, Pasquini MC, MacMillan ML, Wingard JR, Waller EK, Grupp SA, McCarthy PL, Wu J, Hu Z, Carter SL, Horowitz MM, Antin JH. Tacrolimus/sirolimus vs. tacrolimus/methotrexate for graft-vs.-host disease prophylaxis after hla-matched, related donor hematopoietic stem cell transplantation: results of blood and marrow transplant clinical trials network trial 0402. *Blood*. 2012; 120:739–739.
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*. 1999; 94:496–509.
- Gray RJ. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*. 1988; 1141–1154.
- He P, Eriksson F, Scheike TH, Zhang MJ. A proportional hazards regression model for the subdistribution with covariates-adjusted censoring weight for competing risks data. *Scandinavian Journal of Statistics*. 2016; 43:103–122.
- Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. CRC Press; 1999.
- Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*. 2005; 61:223–229. [PubMed: 15737097]
- Kosorok M. Introduction to Empirical Processes and Semiparametric Inference. Springer; New York: 2008.
- Lan KG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663.
- Latouche A, Porcher R, Chevret S. Sample size formula for proportional hazards modelling of competing risks. *Statistics in medicine*. 2004; 23:3263–3274. [PubMed: 15490425]
- Logan BR, Zhang MJ. The use of group sequential designs with common competing risks tests. *Statistics in medicine*. 2013; 32:899–913. [PubMed: 22945865]
- O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 549–556. [PubMed: 497341]
- Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64:191–199.
- Pollard D. Empirical processes: theory and applications. NSF-CBMS regional conference series in probability and statistics. 1990;i–86. JSTOR.
- Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association*. 1997; 92:1342–1350.
- Scheike TH, Zhang MJ, Gerds TA. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*. 2008; 95:205–220.
- Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika*. 1983; 70:315–326.
- Tsiatis AA. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*. 1981; 68:311–315.

Figure 1.

Cumulative type I error rate at each stage: Estimates of the cumulative type I error rate at each stage of a group sequential Gray's test (G) and Fine-Gray test (FG) of treatment effect. Box plots summarize the estimates for each choice of n . Black lines indicate the nominal cumulative type I error rates for the tests.

Figure 2.

Cumulative power at each stage: Estimates of the cumulative power at each stage of a group sequential Gray's test (G) and Fine-Gray test (FG) of treatment effect. Box plots summarize the estimates for each choice of influential covariate strength, ϕ . Black lines indicate the nominal cumulative power rates for the tests.

Figure 3.

Influence of the t statistic of imbalance on conditional power: Estimates of the coefficients of t_{imb} and $|t_{imb}|$ in the conditional power models. Row 1 shows plots of point estimates of the effect of $|t_{imb}|$ from models of conditional power under the null hypothesis for stages 1 and 2, while row 2 shows estimates of the effect of t_{imb} for conditional power under the alternative. Box plots summarize these estimates for each choice of influential covariate strength, ϕ , under each group sequential test considered [Gray's test (G) and the Fine-Gray test (FG) of treatment effect].

Figure 4.

Stopping boundaries and test statistics for the example: Stopping boundaries and test statistics for each stage are shown for the alpha spending functions and analysis methods considered. The respective stopping boundaries for the $\rho = 1$ and $\rho = 3$ alpha spending functions are shown by the dotted and solid lines. Stagewise test statistics for the Fine-Gray model based test and Gray's test are displayed as a plus symbol and a circle, respectively.