Lifetime Data Analysis (2020) 26:603–623
https://doi.org/10.1007/s10985-019-09491-z

Group sequential tests for treatment effect on survival and cumulative incidence at a fixed time point

Michael J. Martens[1] · **Brent R. Logan**[2]

Received: 15 December 2018 / Accepted: 7 November 2019 / Published online: 15 November 2019

Abstract

Medical research frequently involves comparing an event time of interest between treatment groups. Rather than comparing the entire survival or cumulative incidence curves, it is sometimes preferable to evaluate these probabilities at a fixed point in time. Performing a covariate adjusted analysis can improve efficiency, even in randomized clinical trials, but no currently available group sequential test for fixed point analysis provides this adjustment. This paper introduces covariate adjusted group sequential pointwise comparisons of survival and cumulative incidence probabilities. Their test statistics have an asymptotic distribution with independent increments, permitting use of common stopping boundary specification methods. These tests are demonstrated through a redesign of BMT CTN 0402, a clinical trial that evaluated a prophylactic treatment for adverse outcomes following blood and marrow transplantation. A simulation study demonstrates that these tests maintain the type I error rate and power at nominal levels under a variety of settings involving influential covariates.

Keywords Competing risks · Direct binomial regression · Graft versus host disease · Group sequential design · Hematopoietic cell transplantation · Survival analysis

1 Introduction

Time to event outcomes are frequently the main interest of clinical trials. In these studies, investigators may wish to evaluate the treatment effect on the survival or cumulative incidence probability at a fixed time point. This could be preferable to a

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10985-019-09491-z) contains supplementary material, which is available to authorized users.

✉ Michael J. Martens
mmartens@emmes.com

1 The Emmes Company, Rockville, MD 20850, USA

2 Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI 53226, USA

comparison of entire survival or cumulative incidence curves if the particular time point of interest has a strong clinical significance. For example, the Blood and Marrow Transplant Clinical Trials Network (BMT CTN) 0402 (Cutler et al. 2012) was a randomized clinical trial that evaluated an experimental treatment for the prevention of graft versus host disease (GVHD) following blood and marrow transplantation. Its primary endpoint was acute GVHD-free survival at 114 days post-transplant, a time point at which the majority of acute GVHD events were expected to have occurred and any benefit of the new treatment would be most readily observable.

A fixed point comparison may also be favored in the presence of crossing survival or cumulative incidence curves of the treatment groups. This situation may arise in a trial comparing survival rates between surgical and medicinal treatments of malignancy where, due to the risk of complications during and infection after surgery, patients in the surgical arm may have increased risk of mortality immediately following surgery but lower risk thereafter. This complication is illustrated by BMT CTN 1102, an ongoing biologic assignment trial comparing two therapies for MDS: blood and marrow transplantation; and non-transplant, hypomethylating therapy. In cases such as this, common entire curve comparison methods like the log rank test, Gray's test (1988), the Cox proportional hazards model (1972), and the Fine-Gray model (Fine and Gray 1999), because they are optimal only under proportional hazards, may perform poorly because proportionality is violated. A fixed time analysis, on the other hand, is robust to these violations and remains a viable option for treatment comparison, provided an appropriate time point can be identified. For this reason, the treatment benefit on the primary endpoint of BMT CTN 1102, overall survival (OS), will be evaluated by a pointwise comparison of 3-year OS rather than comparing entire OS curves.

Covariate adjustment is a valuable technique that can offer greater efficiency in treatment evaluation (Robinson and Jewell 1991; Pocock et al. 2002; Zhang et al. 2008) and reduction of the influence of other covariates on analysis results. The comparison of 3-year OS in BMT CTN 1102 incorporates covariate adjustment precisely because allocation to treatment arms is performed via biologic assignment, not randomization, and so the chance that covariate imbalance may occur was not guaranteed to be small. The potential for undue covariate influence is also known to exist for randomized studies, since imbalances on covariates may still arise purely by chance (Peto et al. 1976; Gail et al. 1984; Ciolino et al. 2015). Moreover, Hauck et al. (1998) argues that the covariate-specific inference on the treatment effect provided by an adjusted analysis is more relevant in clinical research and more applicable on the patient level than the population-wide inference given by an unadjusted analysis. Group sequential analysis can also improve efficiency of treatment assessment by reducing the sample size and duration for a clinical trial by permitting early stopping for efficacy and futility. Substantial literature exists detailing how to perform group sequential testing in a way that controls the overall type I error rate of a study for a variety of analysis methods, including t-tests, generalized linear models, and the Cox model (Jennison and Turnbull 1999). However, this methodology is quite limited for other time-to-event methods, particularly for fixed time point analysis.

Group sequential tests for fixed point comparison of survival curves can be obtained from the work of Gu and Lai (1991) and Lin et al. (1996), while Logan and Zhang (2013) developed group sequential tests for fixed point comparison of cumulative inci-

dence curves; however, these methods do not account directly for the influence of other covariates. Techniques have been introduced for performing an adjusted comparison of survival and cumulative incidence curves at a fixed time (Klein and Andersen 2005; Zhang et al. 2007), but they have not been studied in the group sequential setting. This manuscript presents group sequential tests of treatment effect on survival and cumulative incidence at a fixed time that adjust for influential covariates.

Section 2 introduces the direct binomial regression model of survival and cumulative incidence probabilities at a fixed point, discusses estimation of its parameters in a group sequential analysis, presents the asymptotic distribution of this sequence of parameter estimates, and proposes a group sequential test for fixed time point analysis based on this finding. Section 3 presents a Cox model stratified toward an adjusted comparison of survival probabilities, gives the asymptotic distribution of its test statistics in a group sequential setting, and offers a group sequential test employing this method. Via an extensive simulation study, Sect. 4 investigates how well the type I error rate and power of these proposed tests adhere to their asymptotic values for realistic sample sizes and compares performance of the proposed, covariate-adjusted methods to unadjusted ones. The proposed methods are applied in a reanalysis of the BMT CTN 0402 trial data in Sect. 5. A discussion of these results in Sect. 6 concludes the paper.

2 Direct binomial regression at a fixed time point

2.1 Basic quantities

The setting considered is a clinical trial in which the primary interest is comparing survival or cumulative incidence probabilities at a prespecified time point s_0 between two study arms. Without loss of generality, we consider modeling the cumulative incidence of cause 1 events; inference on a survival probability is simply a special case where only a single failure type exists. Both the set of patients accrued and the data available on these patients can differ between analyses in the group sequential setting, requiring consideration of two time scales: the calendar time from study opening, denoted by t, and the patients' time on study, denoted by s. Competing risks data is generated for a random sample of patients whose event times may be right-censored due to loss to follow-up. For the i^{th} patient, let T_i denote the event time since enrollment, ϵ_i the event type, C_i the loss to follow-up censoring time since enrollment, τ_i the calendar time of accrual, and $\mathbf{Z}_i$ a vector of covariates of length p.

We assume an upper bound t^* exists on the duration of the study. Let $c^+ = c \vee 0$ denote the positive part of a scalar c. The observed data for patient i at calendar time t is $\left(X_i(t), \Delta_i(t), \Delta_i(t)\epsilon_i, \tau_i, \mathbf{Z}_i'\right)$, where $X_i(t) = T_i \wedge C_i \wedge (t-\tau_i)^+$ is the observed time on study and $\Delta_i(t) = I\left[T_i \leq C_i \wedge (t-\tau_i)^+\right]$ is the event indicator at calendar time t. The event type is only observed if censoring has not occurred, so the observed data includes the product $\Delta_i(t)\epsilon_i$ but not ϵ_i itself. 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 $(t-\tau_i)^+$. We make these assumptions about the data:

1. T_i and C_i are continuous random variables.
2. The $\left(T_i, C_i, \epsilon_i, \tau_i, \mathbf{Z}_i'\right)$ are independently and identically distributed.
3. Independent censoring, i.e. C_i is independent of $\left(T_i, \epsilon_i, \tau_i, \mathbf{Z}_i'\right)$.
4. Independent accrual, i.e. τ_i is independent of $\left(T_i, C_i, \epsilon_i, \mathbf{Z}_i'\right)$.

Assumption (2) implies that the patients' observations form a random sample from the target population. Assumption (3) implies that occurrence of loss to follow-up is unrelated to event occurrence or time of accrual and is a standard assumption made in the derivation of group sequential tests for survival analysis, particularly those utilizing the Kaplan–Meier estimator (Lin et al. 1996), Cox model (Bilias et al. 1997), and Fine-Gray model (Martens and Logan 2018). Assumption (4) implies that the covariates and occurrences of events and loss to follow-up are unrelated to accrual time, assuring that the patients' data distribution is not changing over calendar time.

2.2 The model

Two quantities are usually the main focus of clinical trials involving time to event outcomes: the survival function $S(s) = P(T \geq s)$ and the cumulative incidence function $F_j(s) = P(T \leq s, \epsilon = j)$. To perform a comparison of survival or cumulative incidence at a time point s_0 while adjusting for covariates, we employ the direct binomial regression model of He (2014) of the form

$$h\left[F_1(s_0|\mathbf{Z})\right] = \Lambda_0(s_0) + \boldsymbol{\beta}(s_0)'\mathbf{Z},$$

where $\Lambda_0(s_0)$ and $\boldsymbol{\beta}(s_0)$ are regression parameters, h is a link function, and Z_1 is the treatment indicator, coded as 1 for the active and 0 for the control treatment. This model is an adaptation of the original direct binomial regression model of Scheike et al. (2008), which evaluates the impact of covariates on the entire survival or cumulative incidence curve. Note that the parameters $\Lambda_0(s_0)$ and $\boldsymbol{\beta}(s_0)$ depend on the chosen time point of interest. For the methodology that follows, we assume that this time point s_0 is prespecified and will suppress the argument for these parameters. Our model of interest is

$$h\left[F_1(s_0|\mathbf{Z})\right] = \Lambda_0 + \boldsymbol{\beta}'\mathbf{Z}. \tag{M1}$$

Let $N_i^1(s) = I(T_i \leq s, \epsilon_i = 1)$ denote the counting process of cause 1 events for patient i. $N_i^1(s_0)$ is unbiased for $F_1(s_0|\mathbf{Z}_i)$ conditional on covariates, but is not observed if the patient is censored before s_0. The variable $I[X_i(t) \leq s_0, \Delta_i(t)\epsilon_i = 1] = N_i^1(s_0)\Delta_i(t)$ indicates whether a cause 1 event has been observed for patient i at a calendar time t; but, it is biased for $F_1(s_0|\mathbf{Z}_i)$ in the presence of censoring. However, because $\Delta_i(t) = 1$ implies $\tau_i \leq t$ and we have independent censoring and accrual, the quantity $\tilde{O}_i(t) = N_i^1(s_0)\Delta_i(t)/H(t, T_i)$ is conditionally unbiased for $F_1(s_0|\mathbf{Z}_i)$, where $H(t, s) = P[B_i(t) \geq s|\tau_i \leq t]$ is the "survival function" of the censoring distribution among patients accrued at calendar time t. Note that $H(t, T_i) = E[\Delta_i(t)|\tau_i \leq t]$. Based on this result, Scheike et al. (2008) proposed using inverse probability of censoring weighting (IPCW) to fit the model with observations $O_i(t) =$

$I[X_i(t) \le s_0, \Delta_i(t)\epsilon_i = 1]/\widehat{H}(t, T_i)$ to estimate $F_1(s_0|\mathbf{Z})$ in the fixed sample setting, where $\widehat{H}(t, s)$ is a consistent estimator for $H(t, s)$. We assume $\widehat{H}(t, s) = \Pi_{u \le s}[1 - \sum_{i \in A(t)} I[B_i(t) = u, B_i(t) < T_i]/\sum_{i \in A(t)} I(X_i(t) \ge u)]$, where $A(t) = \{i : \tau_i \le t\}$ identifies the set of patients accrued at calendar time t. $\widehat{H}(t, s)$ is the Kaplan–Meier estimator of $H(t, s)$ and the independent censoring and accrual assumptions imply that $\widehat{H}(t, s)$ is consistent (Lemma 2 in the Online Resource, Section A).

It is important to note that the method of Scheike et al. (2008) infers covariates' effects on the entire cumulative incidence curve F_1, but does not involve sequential analysis. On the other hand, our method evaluates a main covariate's effect on this function at a specific time point s_0 across calendar times t.

2.3 Parameter estimation

The estimating equation used to fit this model is $\mathbf{U}(\boldsymbol{\theta}, t) = \sum_{i=1}^{n} \mathbf{D}_i I(\tau_i \le t)[O_i(t) - F_1(s_0|\mathbf{Z}_i)]$, where $\mathbf{D}_i = \partial F_1(s_0|\mathbf{Z}_i)/\partial\boldsymbol{\theta}$ and $\boldsymbol{\theta} = (\Lambda_0, \boldsymbol{\beta}')'$ is the full vector of model parameters. Let $\boldsymbol{\theta}_0$ denote the true value of this vector. At each calendar time point t, the value can be found that solves $\mathbf{U}(\widehat{\boldsymbol{\theta}}, t) = \mathbf{0}$, giving an estimating process $\widehat{\boldsymbol{\theta}}(t)$ for $\boldsymbol{\theta}_0$. In turn, its second component $\widehat{\beta}_1(t) := \widehat{\theta}_2(t)$ gives an estimating process for β_{10}. Define $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$, and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'$ for $\mathbf{v} \in \mathbb{R}^k$. To draw inference on $\boldsymbol{\theta}_0$, we assume that the following regularity conditions hold:

1. $P\big[(t - \tau_i)^+ \ge s_0\big] > 0$, $P(T_i \ge s_0) > 0$, and $P(C_i \ge s_0) > 0$ for all $t \in [s_0, t^*]$
2. The covariates are nondegenerate and bounded; i.e. there exists $c_z \in \mathbb{R}$ such that $\|\mathbf{Z}_i\|_\infty \le c_z$ for all i
3. h^{-1} exists and is twice continuously differentiable
4. $E\left\{\mathbf{D}_i^{\otimes 2}\left[\frac{N_i^1(s_0)}{H(t_1 \vee t_2, T_i)} - F_1(s_0|\mathbf{Z}_i)^2\right]\right\} - \int_0^{t^*} \frac{E\left\{\mathbf{D}_i[N_i^1(s_0) - N_i^1(u)]\right\}^{\otimes 2} d\Lambda^C(u)}{P(C_i \wedge T_i \ge u)}$ exists for all $t_1, t_2 \in [0, t^*]$
5. There exists $t_* \in [s_0, t^*]$ such that the eigenvalues of $P(\tau \le t)E\big(\mathbf{D}_i^{\otimes 2}\big)$ are bounded below by a constant $r > 0$ for all $t \in [t_*, t^*]$.

Scheike et al. (2008) assumes similar regularity conditions in the derivations for their direct binomial regression model of entire cumulative incidence curves. Condition 1 implies that a positive probability exists of experiencing an event, censoring, or accrual following any calendar time considered. Covariates are usually bounded by physical, temporal, financial, or other constraints, and so condition 2 will hold. Interestingly, the simulation results in Sect. 5 agree with our asymptotic result even though they involved some unbounded, standard normal covariates that violate condition 2, suggesting that it may not be necessary. Condition 3 is satisfied for common choices of link functions, including the identity, log, logit, and complementary log–log transformations. Conditions 4 and 5 guarantee existence of variance functions for the processes $\mathbf{U}(\boldsymbol{\theta}_0, t)$ and $\widehat{\boldsymbol{\theta}}(t)$. For inference on $\boldsymbol{\theta}_0$, we use the asymptotic distribution of $\widehat{\boldsymbol{\theta}}(t)$, given in Theorem 2.

Theorem 1 *Under model (M1), assumptions 1–4, and regularity conditions 1–4,* $\{n^{-1/2}\mathbf{U}(\boldsymbol{\theta}_0, t)\}_{t \in [0,t^*]} \xrightarrow{d} \{\boldsymbol{\xi}(t)\}_{t \in [0,t^*]}$, *where* $\{\boldsymbol{\xi}(t)\}_{t \in [0,t^*]}$ *is a zero mean Gaussian process with continuous sample paths and covariance function*

$$E\left[\boldsymbol{\xi}(t_1)\boldsymbol{\xi}(t_2)'\right] = P(\tau_i \leq t_1 \wedge t_2)\left(E\left\{\mathbf{D}_i^{\otimes 2}\left[\frac{N_i^1(s_0)}{H(t_1 \vee t_2, T_i)} - F_1(s_0|\mathbf{Z}_i)^2\right]\right\} - \int_0^{t^*} \frac{E\left\{\mathbf{D}_i[N_i^1(s_0) - N_i^1(u)]\right\}^{\otimes 2} d\Lambda^C(u)}{P(C_i \wedge T_i \geq u)}\right),$$

where $\Lambda^C(u)$ is the cumulative hazard function of the C_i.

Theorem 2 *Under model (M1), assumptions 1–4, and regularity conditions 1–5, $\{\sqrt{n}[\widehat{\boldsymbol{\theta}(t)} - \boldsymbol{\theta}_0]\}_{t\in[t_*,t^*]} \xrightarrow{d} \{\boldsymbol{\zeta}(t)\}_{t\in[t_*,t^*]}$, where $\{\boldsymbol{\zeta}(t)\}_{t\in[t_*,t^*]}$ is a zero mean Gaussian process with continuous sample paths and covariance function*

$$E\left[\boldsymbol{\zeta}(t_1)\boldsymbol{\zeta}(t_2)'\right] = P(\tau_i \leq t_1 \vee t_2)^{-1} E\left(\mathbf{D}_i^{\otimes 2}\right)^{-1}\left(E\left\{\mathbf{D}_i^{\otimes 2}\left[\frac{N_i^1(s_0)}{H(t_1 \vee t_2, T_i)} - F_1(s_0|\mathbf{Z}_i)^2\right]\right\} - \int_0^{t^*} \frac{E\left\{\mathbf{D}_i[N_i^1(s_0) - N_i^1(u)]\right\}^{\otimes 2} d\Lambda^C(u)}{P(C_i \wedge T_i \geq u)}\right) E\left(\mathbf{D}_i^{\otimes 2}\right)^{-1}$$

Derivations of these results are found in the Online Resource, Section A. The proofs extensively apply the empirical process theory in Pollard (1990) and Kosorok (2008).

2.4 Group sequential test for treatment effect

Group sequential testing allows early stopping of a study for efficacy and/or futility by permitting analysis of the available trial data at two or more interim calendar times during the course of the study, under the reasoning that overwhelming evidence of a treatment benefit, or the complete absence of such evidence, warrants an early declaration of efficacy or futility. The procedure for a two sided group sequential test is formally defined as follows. Consider a group sequential study with K interim analyses, or stages, planned. At the j^{th} interim analysis, performed at calendar time t_j, a standardized test statistic W_j is computed using all available data. At an intermediate analysis $j < K$, the decision rule is to reject H_0 if $|W_j| > a_j$, accept H_0 if $|W_j| \leq b_j$, and continue to stage $j+1$ if neither rejection nor acceptance occurs. Here, the a_j and b_j represent efficacy and futility boundaries at stage j; early stopping for efficacy and futility at this stage can be omitted by setting $a_j = \infty$ and $b_j = -1$, respectively. If the final analysis K is reached, the decision rule is to reject H_0 if $|W_K| > a_K$ and accept H_0 otherwise. Specifying the stopping boundaries a_j, b_j to meet type I error rate and power specifications requires knowing the joint distribution of the test statistics W_j.

A wide array of common methods are known to give test statistics that follow the canonical distribution (Jennison and Turnbull 1999), exactly or asymptotically. The sequence of statistics $(W_1, \ldots, W_K)$ is said to follow the canonical distribution with the information sequence $(I_1, \ldots, I_K)$ for $\boldsymbol{\theta}$ if its asymptotic distribution is multivariate normal with $W_j \sim N(\boldsymbol{\theta}, 1)$ and $Cov(W_i, W_j) = \sqrt{I_i/I_j}$ for $i \leq j$. This distribution is also termed an independent increments structure.

By Theorem 2, the asymptotic covariance of $\widehat{\boldsymbol{\theta}}(t_i)$ and $\widehat{\boldsymbol{\theta}}(t_j)$ depends only on the maximum time of t_i and t_j for any pair $t_i, t_j \in [0, t^*]$. This is also true of the covariance

of the second components, $\widehat{\beta}_1(t_i)$ and $\widehat{\beta}_1(t_j)$. Letting $I_j = E\left[\boldsymbol{\zeta}(t_j)^{\otimes 2}\right]_{22}^{-1}$ be the reciprocal of the (2,2) component of $E\left[\boldsymbol{\zeta}(t_j)^{\otimes 2}\right]$ and $W_j = \widehat{\beta}_1(t_j)\sqrt{I_j}$, it can be seen that $(W_1, \ldots, W_K)$ has independent increments for β_1. Generalized estimating equations (GEE) can be used to obtain estimates and robust standard errors for $\widehat{\beta}_1(t)$ at interim analysis times by supplying the weighted observations $O_i(t)$ as data and specifying the desired link function. This provides a group sequential Wald test of treatment effect at s_0 (i.e., whether $\beta_1 = 0$) by comparing the standardized test statistics W_j to appropriate stopping boundaries.

Pocock and O'Brien-Fleming designs can be employed for boundary specification, but they rely on an assumption of equal increments in information between stages. With time to event outcomes, the observed information levels can deviate greatly from expected levels, especially when sample sizes are not large. Therefore, it is generally preferred to use error spending functions to compute stopping boundaries for each stage using the observed information levels.

3 Treatment comparison by stratification of the Cox model

3.1 The model

The previous section covered the general setting of competing risks data, of which survival data can be considered a special case. This section focuses specifically on the single failure cause/survival data setting and introduces an alternative method for pointwise comparison of survival probabilities via a treatment stratified Cox model (Andersen et al. 1993, Chapter 7). This permits a nonproportional effect of treatment over the time period considered while assuming proportionality of other covariates' effects. If this assumption is correct, this test may be more efficient in detecting a treatment effect than direct binomial regression.

Let $\mathbf{X} = (Z_2, \ldots, Z_p)'$ denote the vector of non-treatment covariates. A Cox model stratified on treatment specifies the treatment-specific hazard rates through the form

$$\lambda(s|Z_1 = k, \mathbf{X}) = \lambda_{k0}(s)\exp(\boldsymbol{\beta}'\mathbf{X}) \text{ for } k = 0, 1 \text{ and } s \leq t^*, \qquad \text{(M2)}$$

where $\lambda_{k0}(s)$ is the baseline hazard function for treatment group $Z_1 = k$. The null hypothesis that no treatment effect on survival at s_0 exists can be stated as $H_0 : S(s_0|Z_1 = 1, \mathbf{X}) = S(s_0|Z_1 = 0, \mathbf{X})$ for all $\mathbf{X}$. Let $\Lambda_{k0}(s_0) = \int_0^{s_0} \lambda_{k0}(u)du$ be the cumulative baseline hazard for strata k. Because a one-to-one relationship exists between the cumulative hazard function and the survival function, an equivalent hypothesis under this model is $H_0 : \Lambda_{10}(s_0) = \Lambda_{00}(s_0)$.

3.2 Estimation

To compare the baseline cumulative hazards at survival time s_0 for the two strata, we employ Breslow's estimators of the strata-specific cumulative hazards. In a sequential analysis, the available trial data is accumulating over calendar time t,

so these estimators will depend on t. Let $\widehat{\Lambda}_{k0}(t, s_0)$ denote Breslow's estimator of the baseline cumulative hazard at survival time s_0 for strata k using all available data at calendar time t. The test statistic used to test H_0 at calendar time t, then, is $A(t) = \widehat{\Lambda}_{10}(t, s_0) - \widehat{\Lambda}_{00}(t, s_0)$; this estimates the difference in cumulative baseline hazards between treatments, $\Lambda_{10}(s_0) - \Lambda_{00}(s_0)$, which is 0 under H_0.

To derive the asymptotic distribution of the test statistic process $A(t)$, we assume the following regularity conditions hold. They are similar to regularity conditions (1)–(5) for the group sequential direct binomial regression test and are stronger than those assumed by Bilias et al. (1997) for a group sequential, unstratified Cox model analysis:

6. $P\left[(t - \tau_i)^+ \geq s\right] > 0$, $P(T_i \geq s) > 0$, and $P(C_i \geq s) > 0$ for all $(t, s) \in D = \left\{(u, v) : 0 \leq v \leq u \leq t^*\right\}$
7. The covariates are bounded; i.e. $\|\mathbf{Z}_i\|_\infty \leq c_z$ for all i
8. There exists $\psi \in (0, 1)$ such that $\lim_{n\to\infty} n_1/n = \psi$, where n_1 is the number of subjects with $Z_1 = 1$;
9. $\boldsymbol{\Gamma}_{kj}(s) = E\left[Y_{ki}^1(s)\mathbf{X}_{ki}^{\otimes j} e^{\boldsymbol{\beta}_0'\mathbf{X}_{ki}}\right]$ exists for all $s \in [0, t^*]$, all $j = 0, 1, 2$, and all $k = 0, 1$
10. There exists $t_* \in (0, t^*)$ such that $\widetilde{\boldsymbol{\Sigma}}(t, s) = E[-\partial \mathbf{U}(\boldsymbol{\beta}, t, s)/\partial \boldsymbol{\beta}']|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is positive definite for $(t, s) \in D_* = \left\{(u, v) : t_* \leq v \leq u \leq t^*\right\}$ with eigenvalues uniformly bounded away from 0.

Conditions (6) and (7) are similar to conditions (1) and (2) assumed for the direct binomial regression model, with similar implications. Condition (8) assumes that the allocation ratio between treatment arms converges to some value in (0,1). Conditions (9) and (10) are similar to regularity conditions 2 and 4 of Bilias et al. (1997) and ensure that $\widehat{\boldsymbol{\beta}}(t)$ and $\widehat{\Lambda}_{k0}(t, s_0)$ converge to Gaussian processes.

The setting considered has a fixed survival time point of interest, s_0, whose cumulative hazard is estimated across calendar times t. Because the Breslow's estimators $\widehat{\Lambda}_{k0}(t, s_0), k = 0, 1$, assess the cumulative hazard at a single survival time, it can be shown that these estimators each have independent increments as the calendar times t vary over $[s_0, t^*]$. Their difference, A(t), also has independent increments; its asymptotic distribution is given by the following theorem.

Theorem 3 *Under model* (M2)*, assumptions (1)–(4), regularity conditions (6)–(10), and* $H_0 : \Lambda_{10}(s_0) = \Lambda_{00}(s_0)$*,* $\{n^{-1/2}A(t)\}_{t\in[t_*,t^*]} \xrightarrow{d} \{v(t)\}_{t\in[t_*,t^*]}$*, where* v *is a zero mean Gaussian process with continuous sample paths and covariance function*

$$E[v(t_1)v(t_2)] = \frac{1}{\psi}\int_0^{s_0} \frac{d\Lambda_{10}(u)}{G(t_1 \vee t_2, u)\Gamma_{10}(u)} + \frac{1}{1-\psi}\int_0^{s_0} \frac{d\Lambda_{00}(u)}{G(t_1 \vee t_2, u)\Gamma_{00}(u)} + (\mathbf{h}_1 - \mathbf{h}_0)'\widetilde{\boldsymbol{\Sigma}}(t_1 \vee t_2, t_1 \vee t_2)^{-1}(\mathbf{h}_1 - \mathbf{h}_0),$$

$\mathbf{h}_k = \int_0^{s_0}\left[\boldsymbol{\Gamma}_{k1}(u)/\Gamma_{k0}(u)\right]d\Lambda_{k0}(u)$*, and* $G(t, s) = P[C_i \wedge (t - \tau_i)^+ \geq s]$ *is the 'survival function' of censoring.*

The proof of this theorem is found in the Online Resource, Section B.

3.3 Group sequential test for treatment effect

Since it depends only on the maximum of the calendar times t_1 and t_2, the process ν also has independent increments. A consistent estimator for the asymptotic variance in Theorem 3 can be obtained by substituting empirical estimators for the quantities in its expression. Thus, a group sequential test of treatment effect at s_0 based on the stratified Cox model is obtained similarly to that for direct binomial regression by comparing the standardized test statistics to the stopping boundaries of the prescribed design.

4 Simulation study

4.1 Design

A simulation study was conducted with two objectives: first, to evaluate how well the proposed tests adhere to the nominal type I error and power specifications for a group sequential trial with realistic sample sizes; and second, to compare the power attained between these methods and common tests that do not perform covariate adjustment. The survival probability at a fixed point was compared between treatment groups using direct binomial regression, the stratified Cox model, and two-sample comparisons of Kaplan–Meier estimators and Nelson–Aalen estimators. Cumulative incidence at a fixed time was compared using direct binomial regression and a two-sample comparison of Aalen–Johansen estimators of cumulative incidence. The two sample comparisons of Kaplan–Meier and Nelson–Aalen estimators were shown to have independent increments by Lin et al. (1996), while Logan and Zhang (2013) showed this result for the comparison of Aalen–Johansen estimators. Parameters of the direct binomial regression tests were estimated using GEE.

We simulated randomized clinical trials comparing efficacy of two treatments through two-sided testing of the existence of a treatment effect. This involved tests of $H_0 : \beta_1 = 0$ for the direct binomial models, $H_0 : \Lambda_{10}(s_0) = \Lambda_{00}(s_0)$ for the stratified Cox model based test, $H_0 : S_1(s_0) = S_0(s_0)$ for the Kaplan–Meier and Nelson–Aalen based tests, and $H_0 : F_{11}(s_0) = F_{10}(s_0)$ for the Aalen–Johansen based test. Treatment assignments were randomized at a 1:1 ratio. Each trial consisted of an accrual period $[0, A]$ during which patients are enrolled in the study at a uniform rate. Group sequential testing was performed at the 0.05 level 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 for each testing method at calendar times where we expect information fractions of 1/3 and 2/3 under the method's assumptions. The total information level and interim analysis times were computed for each combination of simulation settings via Monte Carlo estimation. The complementary log–log link function $x \mapsto \log[-\log(1-x)]$ was used for the direct binomial regression, producing models of the form $S(s_0|\mathbf{Z}) = \exp(e^{\Lambda_0 + \boldsymbol{\beta}'\mathbf{Z}})$ and $F_1(s_0|\mathbf{Z}) = 1 - \exp(e^{\Lambda_0 + \boldsymbol{\beta}'\mathbf{Z}})$ for survival and cumulative incidence at time s_0.

The data were generated according to the following distributions. For the survival model, we assume that event times $T|\mathbf{Z}$ are exponentially distributed with

rate $\exp(\boldsymbol{\beta}'\mathbf{Z})$. In the competing risks model, we generate data from the subdistributions

$$F_1(s|\mathbf{Z}) = 1-[1-q\{1-\exp(-s)\}]^{\exp(\boldsymbol{\beta}'\mathbf{Z})} \quad \text{and} \quad F_2(s|\mathbf{Z})=(1-q)^{\exp(\boldsymbol{\beta}'\mathbf{Z})}(1-e^{-s}).$$

This places a proportional subdistribution hazards model on cause 1 events, where q is a parameter controlling the proportion of cause 1 events observed.

Choices of simulation parameters consisted of the following: sample sizes per treatment group, n = 100, 200, or 400; accrual period A = 2 or 4 years; independent censoring, exponentially distributed at rate of 5% or 10% per year; a treatment effect β_1 of 0 or δ, the value at which the testing method has 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. For the survival models, s_0 was chosen so that $S(s_0) = 0.25, 0.50$, or 0.75, while for competing risks models, we set $s_0 = \log 5$ and chose q so that $F_1(s_0) = 0.2, 0.4$, or 0.6.

We considered two scenarios, each corresponding to a covariate specification. Scenario 1 had one $N(0, 1)$ covariate considered with $\beta_2 = \phi$. Scenario 2 had two covariates, one $N(0, 1)$ and one Bernoulli(0.5), with $\beta_2 = (2/\sqrt{5})\phi$ and $\beta_3 = (1/\sqrt{5})\phi$. The size of covariate effect(s) were selected so that the linear predictor $\boldsymbol{\beta}'\mathbf{Z}$ will have the same mean and variance and, thus, a similar effect on the survival and cumulative incidence probabilities in both scenarios. This gives 144 possible specifications for each model under each value of β_1 considered. We assessed the performance of the proposed tests with regards to type I error rate, power, and conditional power using Monte Carlo estimates obtained from 10,000 simulated trials for each specification.

4.2 Type I error and power

From the simulations, we obtained estimates of the cumulative type I error rates and power over the three stages of the group sequential design for all testing methods considered. Figure 1a, b compares empirical estimates of stagewise type I error rates to their nominal levels for the proposed survival and cumulative incidence tests. To summarize the large number of simulation results, we collect these estimates across choices of s_0, A, censoring rate, and ϕ into box plots that describe the type I error rate for each choice of n; other simulation parameters were found to have no appreciable effect on the type I error rate. These figures show that the direct binomial regression models are conservative at stages 1 and 2, but attain the nominal overall type I error rate. On the other hand, the stratified Cox test is conservative at all stages, though less so for larger samples.

Empirical stagewise estimates of cumulative power for the survival and cumulative incidence tests are shown in Fig. 1c, d. To evaluate how closely the actual power of each testing method comes to the nominal power obtained from its asymptotic distribution, corresponding simulations for each test were performed using a value of δ at which that test has 80% nominal overall power. The stratified Cox test falls below the nominal levels at stages 1 and 2, while the direct binomial test fails to meet nominal levels only

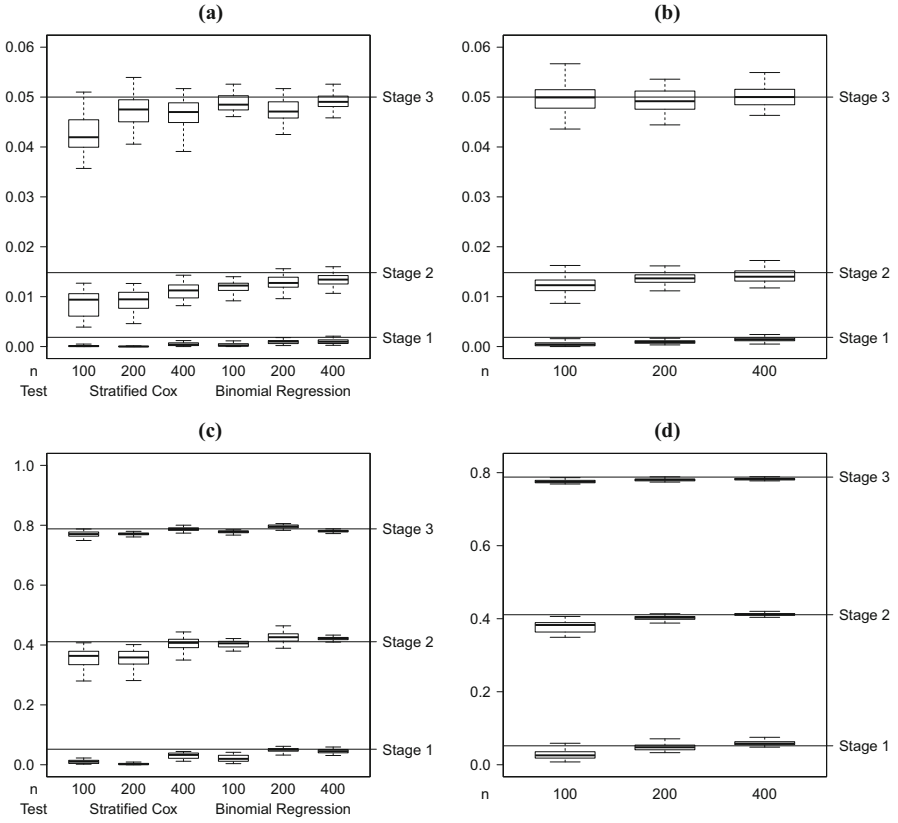

Fig. 1 Stagewise cumulative type I error rate and power: Empirical estimates of the cumulative type I error rate and power for proposed group sequential tests of treatment effect on survival probability are shown in panels (**a**) and (**c**), and for the proposed test on cumulative incidence in panels (**b**) and (**d**). Box plots summarize the estimates for each choice of n. Black lines indicate the nominal cumulative type I error and power levels for the tests

at stage 1 with sample sizes of 100 per group. Both methods attain the targeted overall power level for larger sample sizes. The proposed tests are meeting the specified type I error rate and power requirements for the larger sample sizes considered.

We also performed simulations to compare the power between all methods considered. Figure 2a shows the stagewise cumulative power of tests of treatment effect on fixed time survival from direct binomial regression, a stratified Cox model, and two-sample comparisons of Kaplan–Meier and Nelson–Aalen estimators. The direct binomial regression model is used as the reference test, with δ specified for these simulations as the value at which this test has overall power of 80% under its asymptotic distribution. Estimates are aggregated into box plots for each value of ϕ, the strength of covariate influence. First, the power of the Cox model based test is superior to the direct binomial test by roughly 4–5%. Since the data is generated from a Cox model for scenarios 1 and 2, this is not surprising. Second, neither of the proposed tests are adversely affected by the strength of covariate influence, while the unadjusted methods

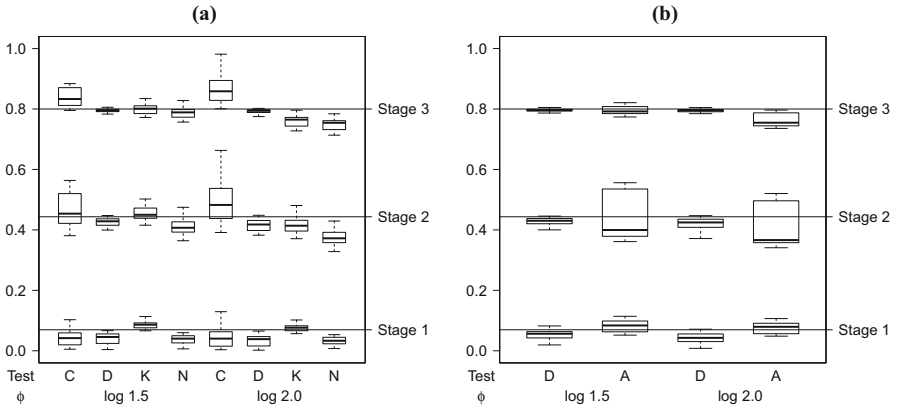

Fig. 2 Comparison of stagewise cumulative power under proportional hazards: Empirical estimates of the cumulative power of group sequential tests of treatment effect, with the effect size δ chosen to provide 80% power under direct binomial regression's asymptotic distribution. Panel (**a**) summarizes estimates for tests on survival probability based on a stratified Cox model (C), direct binomial regression (D), and two-sample comparisons of Kaplan–Meier (K) and Nelson–Aalen (A) estimators, while panel (**b**) presents those for tests on cumulative incidence based on direct binomial regression (D) and a two-sample comparison of Aalen–Johansen (A) estimators. Box plots summarize the estimates for each choice of influential covariate strength, ϕ. Black lines indicate the nominal cumulative power levels

are. Third, these unadjusted tests have an overall power close to 80% for the smaller value of ϕ, while they fall short at the larger value.

Similar plots are shown in Fig. 2b for tests of treatment effect on cumulative incidence from direct binomial regression and a two-sample comparison of Aalen–Johansen estimators. Again, the direct binomial regression based test is the reference, with the treatment effect selected so that this test has 80% nominal power. We see similar trends as with Fig. 2a: the Aalen–Johansen test is sensitive to covariate influence, attaining overall power close to 80% under the smaller choice of ϕ but being underpowered for the larger one. To conduct a suitable comparison of stagewise power levels, the interim analysis times were chosen separately for each method to give expected information fractions of 1/3 and 2/3, provided the assumptions of the method hold. With the two sample comparison of Aalen–Johansen estimators, the assumption is that within each sample, the members are homogeneous; since covariates are at play in the data generating distribution, however, this assumption is violated, causing the observed stagewise information increments to deviate greatly from 1/3 and 2/3 in some scenarios. This is the reason for the increased variability of the Aalen–Johansen estimates at stage 2, shown by the wide interquartile ranges.

4.3 Conditional power

During the course of a trial, conditional power to detect a targeted treatment effect can be described as the probability of rejecting the null hypothesis given the data currently available and that the true treatment effect matches the target. Formally, we define the conditional power to detect $\beta_1 = \gamma$ at interim analysis k as $CP_k(\gamma) = P(\text{reject}$

$H_0|$ data at analysis k, $\beta_1 = \gamma$). A conditional power evaluation is sometimes used to curtail a trial as an alternative to prespecified futility stopping boundaries in a group sequential design, using the justification that low conditional power at interim should indicate a small chance of eventually declaring efficacy and, subsequently, that stopping when conditional power is low avoids "sampling to a foregone conclusion".

Ciolino et al. (2014) demonstrated that, when comparing two groups on a normally distributed outcome that has an influential covariate, the conditional power of a two-sample t-test is sensitive to imbalance on the covariate's distributions in the groups and, moreover, that the conditional power can be strongly impacted even at modest levels of imbalance that arise surprisingly often in randomized studies. Thus, our simulation study assessed whether the conditional power of each fixed point method accurately predicts the rejection probability in the presence of influential covariates. We used the t statistic of imbalance (Ciolino et al. 2015) to measure the disparity in covariate distributions between treatment groups for scenario 1. It has the form $t_{imb} = (\bar{Z}_{2,1} - \bar{Z}_{2,0})/(s_z\sqrt{1/n_1 + 1/n_0})$, where s_z is the pooled standard deviation of Z_2, and n_k and $\bar{Z}_{2,k}$ are the number of subjects and the average value of Z_2 in treatment group k. The unconditional power for the test, $P_{\beta_1}(\text{reject } H_0) = E[CP_k(\beta_1)]$, is obtained by integrating over the data's distribution.

To determine whether covariate imbalance impacts the rejection probability, we fit the logistic regression models

$$\text{logit } P_{\beta_1}(\text{reject } H_0 \mid t_{imb,k}) = \text{logit } CP_k(\beta_1) + \nu_1 t_{imb,k} + \nu_2 |t_{imb,k}|$$

for stages $k = 1, 2$ and for $\beta_1 = 0, \delta$, where $t_{imb,k}$ denotes the t statistic of imbalance in the set of patients accrued by interim analysis k. Box plots of the parameter estimates $\hat{\nu}_1$ and $\hat{\nu}_2$ summarize the results of these models under the various choices of simulation parameters for each test considered. The imbalance impacts the conditional likelihood of rejecting the null hypothesis if and only if $\hat{\nu}_1$ and/or $\hat{\nu}_2$ differ from 0 substantially.

Coefficient estimates of $|t_{imb}|$ for the survival tests under the null hypothesis are shown in Fig. 3a, b for stages 1 and 2, respectively. The box plots for the direct binomial and stratified Cox based tests are centered near 0, whereas those for the unadjusted tests show a positive bias that increases as $|t_{imb}|$ does. This implies that imbalance on an influential covariate between treatment groups will cause the conditional power of the unadjusted tests at $\beta_1 = 0$ to underestimate the true type I error rate with increasing severity as the size of the imbalance grows. Moreover, increasing the sample size is not a remedy, as the magnitude of this underestimation is similar for the three sample sizes considered.

Figure 3c, d identify 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 tests but has a negative bias for the others. When a negative imbalance exists between groups with respect to their covariate distributions, the conditional power of the Kaplan–Meier and Nelson–Aalen based tests at $\beta_1 = \delta$ will tend to underestimate the true power of this test; moreover, this inaccuracy will increase as t_{imb} does. These results imply that if low conditional power is used as a basis for early stopping, the trial will be more likely to stop wrongly for futility when a treatment effect actually exists.

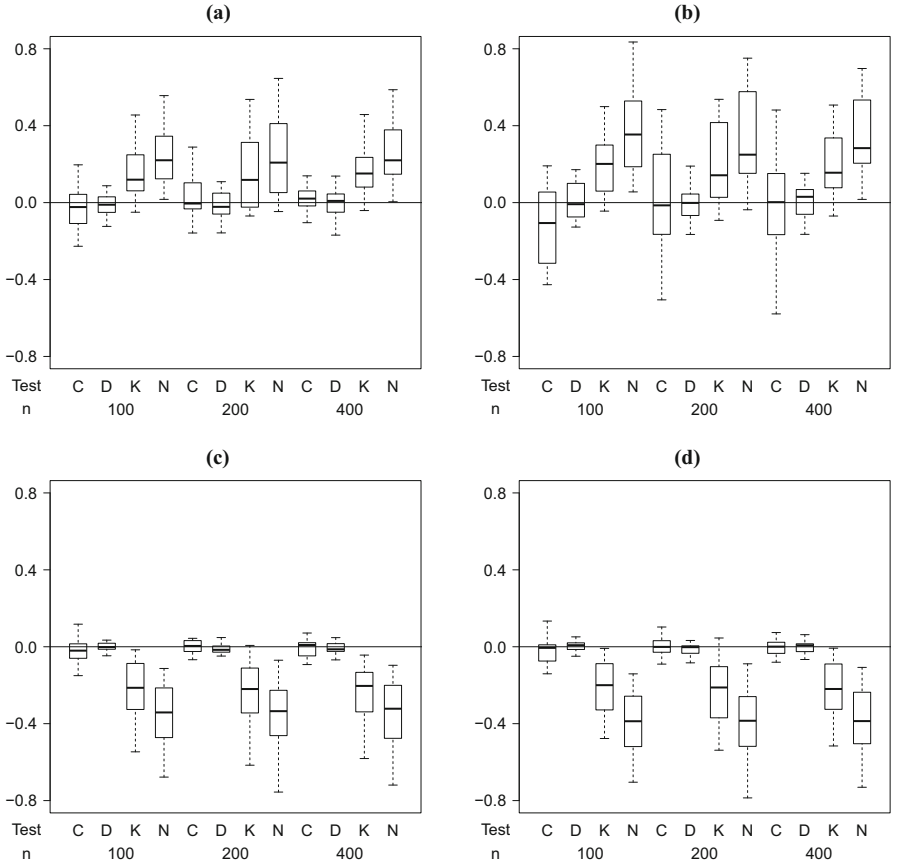

Fig. 3 Influence of the t statistic of imbalance on conditional power: Estimates of the coefficients of t_{imb} and $|t_{imb}|$ in the conditional power models. Row 1 shows plots of point estimates of the effect of $|t_{imb}|$ from models of conditional power under the null hypothesis for stages 1 and 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 sample size, under each group sequential test considered [Stratified Cox (C), direct binomial regression (D), Kaplan–Meier (K), and Nelson–Aalen (N)]

Similar trends are apparent for the competing risks methods considered, illustrated by Fig. 4. Conditional power of the Aalen–Johansen based test fails to represent accurately the rejection probability under either hypothesis, while that of the direct binomial test is valid. Though our simulations did not include futility stopping boundaries in the group sequential design, the implications of covariate imbalance are similar whether prespecified stopping boundaries or conditional power is used for curtailment. Namely, for the unadjusted tests considered, the stagewise type I error rates increase as the size of the covariate imbalance does, while the stagewise power levels decrease as the t statistic of imbalance increases, increasingly favoring the control group.

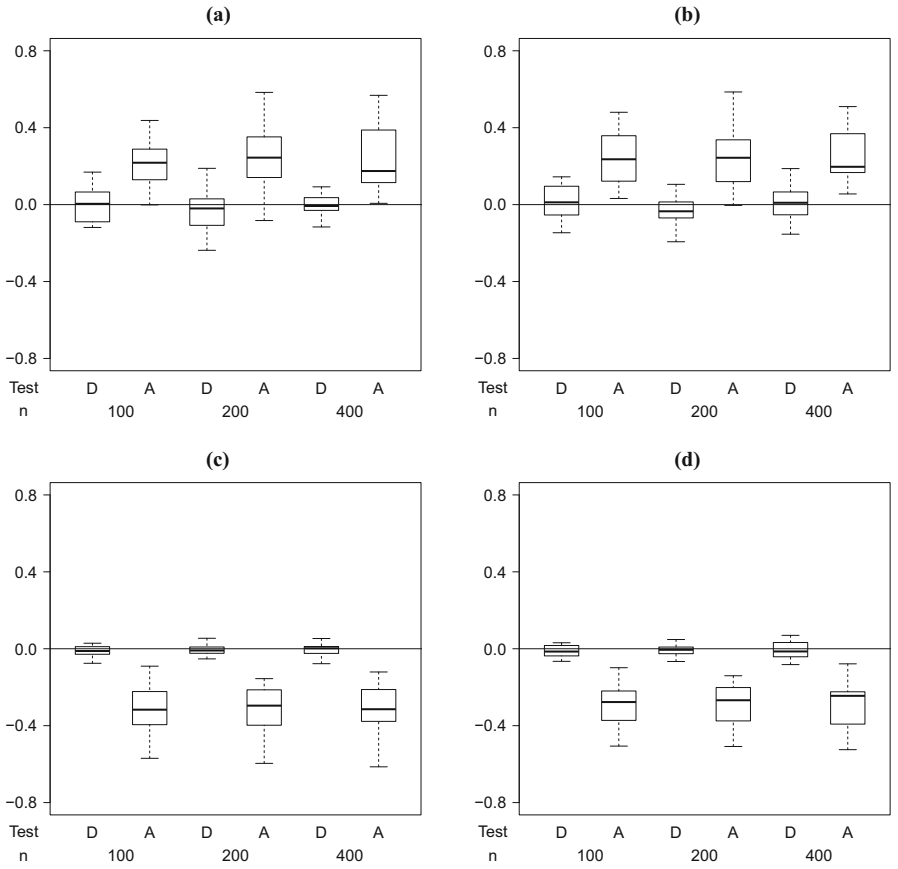

Fig. 4 Influence of the t statistic of imbalance on conditional power: Estimates of the coefficients of t_{imb} and $|t_{imb}|$ in the conditional power models. Row 1 shows plots of point estimates of the effect of $|t_{imb}|$ from models of conditional power under the null hypothesis for stages 1 and 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 sample size, under each group sequential test considered [Direct binomial regression (D) and Aalen–Johansen (A)]

4.4 Nonproportional hazards

Finally, we investigate how robust the survival methods are to deviations from the proportional hazards assumption. This involves a third set of simulations under which survival times are generated as $T|Z, X \sim$ Weibull (shape α, rate γ), where Z is a binary treatment indicator, X is a binary influential covariate, $\alpha = \alpha_0 + aZ + bX$ and $\gamma = \gamma_0 \exp(cZ + dX)$ such that the corresponding survival function has the form $S(t|Z, X) = \exp(-\gamma t^{\alpha})$. The survival function at s_0 can be written in the form of the direct binomial regression model with a complementary log–log link as $S(s_0|Z, X) = \exp[-\exp\{\Lambda_0 + \boldsymbol{\beta}'(Z, X)\}]$, where $\Lambda_0 = \alpha_0 \log s_0 + \log \gamma_0$, $\beta_1 = a \log s_0 + c$, and $\beta_2 = b \log s_0 + d$. Thus, a test of $H_0 : \beta_1 = 0$ is a valid test of whether a treatment effect on survival at s_0 exists.

Furthermore, if X has an effect on the shape parameter, i.e. b is nonzero, hazards will fail to be proportional within treatment groups and the stratified Cox model is violated. To control the degree of this violation, we examine the log hazard ratio of X, $\beta_2 = b \log s_0 + d$. The first summand depends on the survival time considered, while the second does not. Then the ratio $r = b \log s_0 / \beta_2$ represents the proportion of the effect of X that is time varying. We suspect that the performance of the stratified Cox model based test will degrade as this ratio increases from 0 to 1.

For this third scenario, we consider testing for a treatment effect in the presence of crossing survival curves by comparing survival at a late time point, $s_0 = 1.5$ years. Simulations were conducted under these choices of parameters: sample sizes per treatment group, n = 100, 200, or 400; accrual period of 2 or 4 years; independent censoring, exponentially distributed at a rate of 5% or 10% per year; $a = 0$ or δ, a treatment effect at which the proposed test should have 80% power under its test statistics' asymptotic distribution; $\beta_2 = \log 1.5$ or $\log 2$, and $r = 0, 0.25, 0.5$, and 0.75. Other parameters of the data generating distribution are set at $\alpha_0 = 3$, $\lambda_0 = 0.2$, and $c = 0$. This implies that $a = \beta_1 / \log s_0$, so an appropriate value of a can be chosen so that the direct binomial regression model will have 80% nominal power.

Figure 5 shows box plots of cumulative power of the four survival tests for each level of covariate strength and ratio of time-varying effect of the covariate considered. As before, direct binomial regression is used as the reference, with the treatment effect size a chosen under the alternative to give this test a nominal power of 80%. Several trends are apparent. For $r = 0$, where the data are generated from a stratified Cox model, the Cox model based test has superior overall power compared to direct binomial regression; when $r = 0.25$, the Cox based test has slightly higher power; and with $r = 0.5$ and 0.75, direct binomial regression has better power. As the covariate effect β_2 increases, we see a large drop in the power of the Cox model based tests and the unadjusted tests, but no effect on the power for direct binomial regression. The unadjusted tests have overall power near 80% for the smaller value of β_2 but fall short for the larger value. Somewhat surprisingly, the Cox based test has inferior power to the unadjusted tests when the covariate has a strong time varying effect ($r = 0.5$ and 0.75). In summary, the direct binomial regression test is much more robust to the presence of nonproportional hazards and covariate influence than the Cox based and unadjusted tests.

5 Example

We illustrate the application of the proposed tests using data from the Blood and Marrow Transplant Clinical Trials Network 0402 (Cutler et al. 2012), a randomized clinical trial comparing an experimental prophylactic therapy for graft versus host disease (GVHD), Tacrolimus and Sirolimus (Tac/Sir), to the current standard regimen of Tacrolimus and Methotrexate (Tac/Mtx). 304 patients were randomized at a 1:1 ratio to the two arms. The treatment groups did not differ significantly on the primary endpoint, acute GVHD-free survival at day 114 post transplant, a composite endpoint indicating that a patient neither suffered acute GVHD nor died. However, this was evaluated by a pointwise comparison of Kaplan–Meier estimators, which did

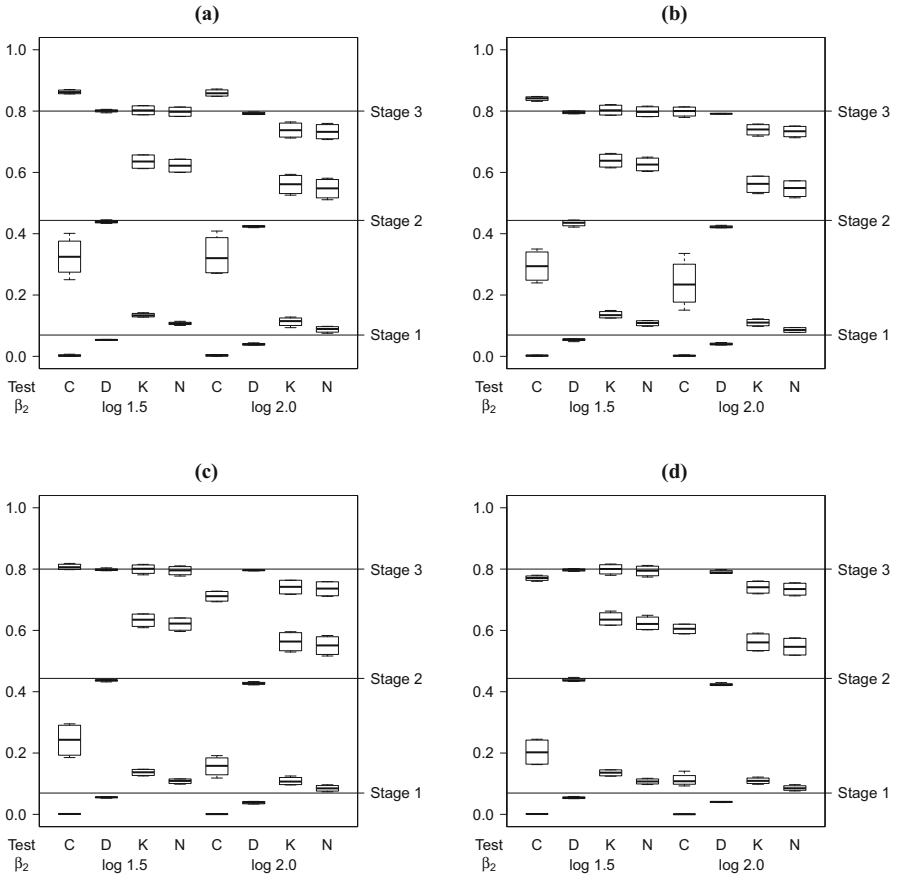

Fig. 5 Comparison of stagewise cumulative power for survival tests under nonproportional hazards: Empirical estimates of the cumulative power at each stage of group sequential tests of treatment effect based on a stratified Cox model (C), direct binomial regression (D), and two-sample comparisons of Kaplan–Meier (K) and Nelson–Aalen (A) estimators. Box plots summarize the estimates for each choice of influential covariate strength, β_2. Plots (**a**)–(**d**) show results of simulations with ratio $r = 0, 0.25, 0.50$, and 0.75, respectively. Black lines indicate the nominal cumulative power rates for the direct binomial regression based test

not account for covariates. Moreover, at 304 patients, this trial is considered a fairly large trial in the field of blood and marrow transplantation; although a fixed sample design was employed for its design and analysis, a group sequential design could have offered the possibility of early stopping and/or early reporting of study results had overwhelming evidence of efficacy or futility been seen at interim.

Our reanalysis of the trial data assesses the treatment effect on two endpoints: acute GVHD-free survival at day 114, a survival outcome, and extensive chronic GHVD (ECGVHD) at 2 years, a competing risk outcome that has death as the competing event. 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 both GVHD and death. The proposed

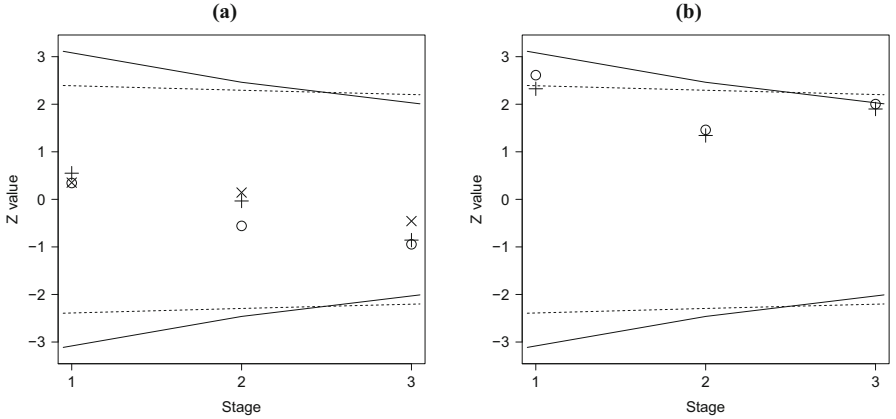

Fig. 6 Group sequential analysis for the examples: Stopping boundaries and test statistics for each stage are shown for the alpha spending functions and analysis methods considered for (**a**) survival tests on acute GVHD-free survival at Day 114 and (**b**) cumulative incidence tests on extensive chronic GVHD at 2 years post-transplant. 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 direct binomial model based tests are shown by a plus symbol, for the Cox model based test by an x symbol, and for the Kaplan–Meier and Aalen–Johansen tests by a circle

tests will be applied to account for the potential influence of these covariates on acute GVHD-free survival and ECGVHD.

A three stage group sequential design is used for treatment evaluation with interim analyses performed at calendar times corresponding to maximum information fractions of 1/3 and 2/3. A type I error spending function is used to permit early stopping for efficacy, selected from the power family in Jennison and Turnbull (1999) with $\rho = 1$ or 3. Direct binomial regression, the stratified Cox model test, and a comparison of Kaplan–Meier estimators were used to test for treatment effect on aGVHD-free survival at day 114. Standardized test statistics for the three stages are $Z_{d1} = 0.550$, $Z_{d2} = -0.334$, and $Z_{d3} = -0.857$ for direct binomial regression; $Z_{c1} = 0.352$, $Z_{c2} = 0.143$, and $Z_{c3} = -0.457$ for the Cox model based test; and $Z_{k1} = 0.344$, $Z_{k2} = -0.558$, and $Z_{k3} = -0.945$ for the Kaplan–Meier based test.

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 $d_3 = 2.009$. With either spending function, all three tests fail to find compelling evidence of a treatment effect on acute GVHD-free survival at day 114 (see Fig. 6a).

On the other hand, for a comparison of treatment effect on the cumulative incidence of ECGVHD at 2 years, the adjusted and unadjusted methods draw different conclusions. Standardized test statistics at the three stages are $Z_{b1} = 2.325$, $Z_{b2} = 1.343$, and $Z_{b3} = 1.900$ for direct binomial regression and are $Z_{a1} = 2.612$, $Z_{a2} = 1.461$, and $Z_{a3} = 2.003$ for a two sample comparison of Aalen–Johansen estimators. If the $\rho = 1$ function is used, the Aalen–Johansen based test will reject the null hypothesis at stage 1, while the binomial regression test will accept the null. Under the $\rho = 3$ function, both tests will accept the null. With $\rho = 1$, the Aalen–Johansen test asserts

that Tac/Sir is inferior to Tac/Mtx in the prevention of ECGVHD, while the direct binomial regression test finds no evidence that the effects differ appreciably. Under the $\rho = 3$ function, neither method finds a treatment effect on ECGVHD occurrence at 2 years (see Fig. 6b).

6 Discussion

This paper introduced group sequential tests for treatment effect on survival and cumulative incidence probabilities at a fixed time that adjust properly for influential covariates, which was verified through a simulation study of clinical trials under realistic conditions. Moreover, these methods offer test statistics whose asymptotic distributions have an independent increments structure, making standard boundary specification methods applicable. Among the two proposed survival tests, the direct binomial regression method offers more robust treatment evaluation than the stratified Cox model based test when the proportionality assumption is violated for covariates. On the other hand, the latter method can provide a modest boost in power when this assumption holds.

The derivations for the direct binomial regression model rely on the assumption of a censoring distribution that is independent of covariates. For some studies, though, this assumption may be untenable. The inverse weighting used in parameter estimation of this model involves an estimator of the censoring distribution function. We may be able to maintain validity of this group sequential method under covariate-dependent censoring through an appropriate adjustment of the censoring distribution's estimator in a manner similar to that employed by He (2014). This requires showing that the estimating process of the censoring distribution converges uniformly in probability to the true distribution. The Kaplan–Meier estimator has this property (Andersen et al. 1993); it may be possible to show that a distributional estimate obtained from the Cox model is also uniformly consistent.

Our methods involve comparing efficacy on survival or cumulative incidence at a single time point using independent subjects. These methods could be extended to accommodate two situations of increased complexity: treatment evaluation at multiple time points, and clustering of subjects. The former case may arise if investigators wish to consider the treatment effect on a collection of time points of clinical significance, while the latter can occur in a multicenter clinical trial. Direct binomial regression at multiple time points with independent subjects was investigated by He (2014), with each time point corresponding to one parameter in the regression model and GEE used to account for within subject correlation across time points. This could be extended to provide a group sequential Wald test of treatment effect at several time points. Similarly, we may be able to derive the asymptotic joint distribution of the Breslow estimator at multiple event times for the stratified Cox model, which would provide a means for an overall test of treatment effect over several time points as well. For analyzing clustered time to event observations, Logan et al. (2011) presented a method based on pseudovalues that fits a marginal regression model using a GEE approach to adjust for within cluster correlation. A similar approach might be applied for parameter

estimation of a marginal direct binomial regression model in both the fixed sample and group sequential settings.

Although inverse probability weighting methods are known for their robustness, their estimates can suffer from high variance. For simplicity, we used identity weights in the estimating equations of the direct binomial regression models. However, this practice is known to be inefficient. A procedure to specify weights accurately for more efficient parameter estimation may enhance the treatment evaluation of this method.

Acknowledgements Support for this study was provided by Grant #F31HL134317 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 Blood and Marrow Transplant Clinical Trials Network for permitting use of the 0402 trial data. The content is solely the responsibility of the authors and does not necessarily represent the official views of the above mentioned parties.

References

Andersen P, Borgan O, Gill R, Keiding N (1993) Statistical models based on counting processes. Springer, Berlin

Bilias Y, Gu M, Ying Z (1997) Towards a general asymptotic theory for Cox model with staggered entry. Ann Stat 25(2):662–682

Ciolino JD, Martin R, Zhao W, Jauch EC, Hill MD, Palesch YY (2014) Continuous covariate imbalance and conditional power for clinical trial interim analyses. Contemp Clin Trials 38(1):9–18

Ciolino JD, Martin RH, Zhao W, Hill MD, Jauch EC, Palesch YY (2015) Measuring continuous baseline covariate imbalances in clinical trial data. Stat Methods Med Res 24(2):255–272

Cox D (1972) Regression and life tables (with discussion). J R Stat Soc Ser B 34(2):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 (2012) 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 120(21):739–739

Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94(446):496–509

Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71(3):431–444

Gray RJ (1988) A class of k-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 16(3):1141–1154

Gu MG, Lai TL (1991) Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. Ann Stat 19(3):1403–1433

Hauck WW, Anderson S, Marcus SM (1998) Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials 19(3):249–256

He P (2014) Bias reduction by using covariate-adjusted censoring weights for survival and competing risks data. Doctoral Dissertation, Medical College of Wisconsin

Jennison C, Turnbull BW (1999) Group sequential methods with applications to clinical trials. CRC Press, Boca Raton

Klein JP, Andersen PK (2005) Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 61(1):223–229

Kosorok M (2008) Introduction to empirical processes and semiparametric inference. Springer, New York

Lin D, Shen L, Ying Z, Breslow N (1996) Group sequential designs for monitoring survival probabilities. Biometrics 52(3):1033–1041

Logan BR, Zhang MJ (2013) The use of group sequential designs with common competing risks tests. Stat Med 32(6):899–913

Logan BR, Zhang M, Klein JP (2011) Marginal models for clustered time-to-event data with competing risks using pseudovalues. Biometrics 67(1):1–7

Martens MJ, Logan BR (2018) A group sequential test for treatment effect based on the Fine-Gray model. Biometrics 74(3):1006–1013

Peto R, Pike M, Armitage P, Breslow N, Cox D, Howard SV, Mantel N, McPherson K, Peto J, Smith P (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. i. Introduction and design. Br J Cancer 34(6):585

Pocock SJ, Assmann SE, Enos LE, Kasten LE (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 21(19):2917–2930

Pollard D (1990) Empirical processes: theory and applications. In: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, Hayward, California; American Statistical Association, Alexandria, Virginia, pp. 1–86

Robinson LD, Jewell NP (1991) Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev 59(2):227–240

Scheike TH, Zhang MJ, Gerds TA (2008) Predicting cumulative incidence probability by direct binomial regression. Biometrika 95(1):205–220

Zhang M, Tsiatis AA, Davidian M (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. Biometrics 64(3):707–715

Zhang X, Loberiza FR, Klein JP, Zhang MJ (2007) A SAS macro for estimation of direct adjusted survival curves based on a stratified Cox regression model. Comput Methods Programs Biomed 88(2):95–101