Biometrika (2024), **111**, 2, *pp.* 691–705
https://doi.org/10.1093/biomet/asad045
Advance Access publication 27 July 2023

Covariate-adjusted log-rank test: guaranteed efficiency gain and universal applicability

By TING YE

Department of Biostatistics, University of Washington,
Box 351617, Seattle, Washington 98195, U.S.A.
tingye1@uw.edu

JUN SHAO

Department of Statistics, University of Wisconsin,
1300 University Avenue, Madison, Wisconsin 53706, U.S.A.
shao@stat.wisc.edu

and YANYAO YI

Global Statistical Sciences, Eli Lilly and Company,
839 S Delaware St., Indianapolis, Indiana 46285, U.S.A.
yi_yanyao@lilly.com

Summary

Nonparametric covariate adjustment is considered for log-rank-type tests of the treatment effect with right-censored time-to-event data from clinical trials applying covariate-adaptive randomization. Our proposed covariate-adjusted log-rank test has a simple explicit formula and a guaranteed efficiency gain over the unadjusted test. We also show that our proposed test achieves universal applicability in the sense that the same formula of test can be universally applied to simple randomization and all commonly used covariate-adaptive randomization schemes such as the stratified permuted block and the Pocock–Simon minimization, which is not a property enjoyed by the unadjusted log-rank test. Our method is supported by novel asymptotic theory and empirical results for Type-I error and power of tests.

Some key words: Covariate calibration; Minimization; Permuted block; Pitman's relative efficiency; Stratification; Time-to-event data; Validity and power of tests.

1. Introduction

In clinical trials, adjusting for baseline covariates has been widely advocated as a way to improve efficiency for demonstrating treatment effects 'under approximately the same minimal statistical assumptions that would be needed for unadjusted estimation' (ICH E9, 1998; EMA, 2015; FDA, 2023). In testing for an effect between two treatments with right-censored time-to-event outcomes, adjusting for covariates using the Cox proportional hazards model has been demonstrated to yield valid tests even if the Cox model is misspecified (Lin & Wei, 1989; Kong & Slud, 1997; DiRienzo & Lagakos, 2002). However, these tests may be less powerful than the log-rank test that does not adjust for any covariates when the Cox model is misspecified (Kong & Slud, 1997). Although efforts have been made to improve the efficiency of the log-rank test through covariate adjustment from semiparametric theory (Lu & Tsiatis, 2008; Moore & van der Laan, 2009), the solutions are complicated and their validity is established only under simple randomization, i.e., treatments are assigned to patients completely at random.

To balance the number of patients in each treatment arm across baseline prognostic factors in clinical trials with sequentially arrived patients, covariate-adaptive randomization has become the new norm. From 1989 to 2008, covariate-adaptive randomization was used in more than 500 clinical trials (Taves, 2010); among nearly 300 trials published in two years, 2009 and 2014, 237 of them applied covariate-adaptive randomization (Ciolino et al., 2019). The two most popular covariate-adaptive randomization schemes are the stratified permuted block (Zelen, 1974) and the Pocock–Simon minimization (Taves, 1974; Pocock & Simon, 1975). Other schemes can be found in the reviews of Schulz & Grimes (2002) and Shao (2021). Unlike simple randomization, covariate-adaptive randomization generates a dependent sequence of treatment assignments, which may render conventional methods developed under simple randomization not necessarily valid under covariate-adaptive randomization (EMA, 2015; FDA, 2023). For time-to-event data under covariate-adaptive randomization, Ye & Shao (2020) showed that some conventional tests including the log-rank test are conservative and Wang et al. (2023) showed that the Kaplan–Meier estimator of the survival function has reduced variance compared to that under simple randomization.

The discussion so far has brought up two issues in adjusting for covariates. First is the need for guaranteed efficiency gains over unadjusted methods, without requiring additional assumptions. Second is the need for methods with wide applicability to all commonly used covariate-adaptive randomizations. These issues have been well addressed when adjustments are made under linear working models for non-time-to-event data (Tsiatis et al., 2008; Zhang et al., 2008; Lin, 2013; Ye et al., 2022). Ye et al. (2022) also showed that adjustment via linear working models can achieve universal applicability in the sense that the same inference procedure can be universally applied to all commonly used covariate-adaptive randomization schemes, a desirable property for application. For right-censored time-to-event outcomes, to the best of our knowledge, no result has been established for covariate adjustment with guaranteed efficiency gain and universal applicability.

In this paper we propose a nonparametric covariate adjustment method for the log-rank test, which has a simple explicit form and can achieve the goal of guaranteed efficiency gain over the unadjusted log-rank test as well as universal applicability to simple randomization and all commonly used covariate-adaptive randomization schemes. The unadjusted log-rank test is not valid under covariate-adaptive randomization; although it can be modified to be applicable to some randomization schemes (Ye & Shao, 2020), the modification needs to be tailored to each randomization scheme, i.e., no universal applicability. Our main idea is to obtain a particular derived outcome for each patient from linearizing the log-rank test statistic and then apply the generalized regression adjustment or augmentation (Cassel et al., 1976; Lu & Tsiatis, 2008; Tsiatis et al., 2008; Zhang et al., 2008) to the derived outcomes. We also develop parallel results for the stratified log-rank test with adjustment for additional covariates. Our proposed tests are supported by novel asymptotic theory of the existing and proposed statistics under the null hypothesis and alternative without requiring any specific model assumption, and under all commonly used covariate-adaptive randomization schemes. Estimation and confidence intervals for treatment effects after testing are also discussed. Our theoretical results are corroborated by a simulation study that examines finite sample Type-I error and power of tests. A real data example is included for illustration.

2. Preliminaries

For a patient from the population under investigation, let T_j and C_j be the potential failure time and right-censoring time, respectively, under treatment $j = 0$ or 1, and W be a vector containing all observed baseline covariates. Suppose that a random sample of n patients is obtained from the population with independent $(T_{i0}, C_{i0}, T_{i1}, C_{i1}, W_i)$ $(i = 1, \ldots, n)$, identically distributed as (T_0, C_0, T_1, C_1, W). For each patient, only one of the two treatments is received. Thus, if patient i receives treatment j, then the observed outcome with possible right censoring is $\{\min(T_{ij}, C_{ij}), \delta_{ij}\}$, where δ_{ij} is the indicator of the event $T_{ij} \leqslant C_{ij}$.

Let I_i be a binary treatment indicator for patient i and $0 < \pi < 1$ be the prespecified treatment assignment proportion for treatment 1. Consider the design, i.e., the generation of the I_i for n sequentially arrived patients. Simple randomization assigns patients to treatments completely at random with $\mathrm{pr}(I_i = 1) = \pi$ for all i, which does not make use of baseline covariates and may yield treatment proportions that substantially deviate from the target π across levels of some prognostic factors. Because of this, covariate-adaptive randomization using a subvector Z of W is widely applied, which does not use any model and is nonparametric. All commonly used covariate-adaptive randomization schemes satisfy the following mild condition (Baldi Antognini & Zagoraiou, 2015).

Condition 1. The covariate Z for which we want to balance in treatment assignment is an observed discrete baseline covariate with finitely many joint levels; conditioned on $(Z_i, i = 1, \ldots, n)$, $(I_i, i = 1, \ldots, n)$ is conditionally independent of $(T_{i1}, C_{i1}, T_{i0}, C_{i0}, W_i, i = 1, \ldots, n)$; $E(I_i \mid Z_1, \ldots, Z_n) = \pi$ for all i and, for every level z of Z, $n_{z1}/n_z \to \pi$ in probability as $n \to \infty$, where n_z is the number of patients with $Z_i = z$ and n_{z1} is the number of patients with $Z_i = z$ and $I_i = 1$.

Although simple randomization is not counted as covariate-adaptive randomization, it also satisfies Condition 1.

We focus on testing the following null hypothesis of the no-treatment effect, which is the null hypothesis when the conventional log-rank test is applied: $H_0\colon \lambda_1(t) = \lambda_0(t)$ for all times t, versus the alternative that H_0 does not hold, where $\lambda_j(t)$ is the unspecified hazard function of T_j, unconditional on covariates.

After data are collected from all patients, a test statistic $\mathcal{T}$ is a function of observed data, constructed such that H_0 is rejected if and only if $|\mathcal{T}| > z_{\alpha/2}$, where α is a given significance level and $z_{\alpha/2}$ is the $(1 - \alpha/2)$th quantile of the standard normal distribution. A test $\mathcal{T}$ is asymptotically valid if, under H_0, $\lim_{n\to\infty} \mathrm{pr}(|\mathcal{T}| > z_{\alpha/2}) \leqslant \alpha$, with equality holding for at least one parameter value under the null hypothesis H_0. A test $\mathcal{T}$ is asymptotically conservative if, under H_0, there exists an α_0 such that $\lim_{n\to\infty} \mathrm{pr}(|\mathcal{T}| > z_{\alpha/2}) \leqslant \alpha_0 < \alpha$.

The test statistic of the log-rank test is

$$\mathcal{T}_{\mathrm{L}} = n^{1/2}\hat{U}_{\mathrm{L}}/\hat{\sigma}_{\mathrm{L}} \tag{1}$$

(Mantel, 1966; Kalbfleisch & Prentice, 2011), where

$$\hat{U}_{\mathrm{L}} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{I_i - \frac{\bar{Y}_1(t)}{\bar{Y}(t)}\right\}\mathrm{d}N_i(t), \qquad \hat{\sigma}^2_{\mathrm{L}} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\frac{\bar{Y}_1(t)\,\bar{Y}_0(t)}{\bar{Y}(t)^2}\,\mathrm{d}N_i(t),$$

$$\bar{Y}_1(t) = \sum_{i=1}^{n}\frac{I_i\,Y_i(t)}{n}, \qquad \bar{Y}_0(t) = \sum_{i=1}^{n}\frac{(1-I_i)\,Y_i(t)}{n},$$

$$\bar{Y}(t) = \bar{Y}_1(t) + \bar{Y}_0(t) \quad \text{and} \quad Y_i(t) = I_i\,Y_{i1}(t) + (1-I_i)\,Y_{i0}(t),$$

with $Y_{ij}(t)$ the indicator of the event $\min(T_{ij}, C_{ij}) \geqslant t$, $N_i(t) = I_i N_{i1}(t) + (1-I_i)N_{i0}(t)$ the counting process of observed failures, $N_{ij}(t)$ the indicator of the event $T_{ij} \leqslant \min(t, C_{ij})$, and the upper limit τ in the integral a point satisfying $\mathrm{pr}\{\min(T_{ij}, C_{ij}) \geqslant \tau\} > 0$ for $j = 0, 1$.

The log-rank test $\mathcal{T}_{\mathrm{L}}$ in (1) is valid under simple randomization and the following assumption.

Assumption 1. We have $C_I \perp T_I \mid I$, where I is the treatment indicator, $\perp$ denotes independence and the vertical line denotes conditioning.

Assumption 1 is needed for a valid nonparametric log-rank test without requiring any model on T_j or C_j (Kong & Slud, 1997; DiRienzo & Lagakos, 2002; Lu & Tsiatis, 2008; Parast et al., 2014; Zhang, 2015).

As $\mathcal{T}_{\mathrm{L}}$ does not utilize any baseline covariate information, it is used as the benchmark in considering baseline covariate adjustment for efficiency gain, under the same Assumption 1 'that would be needed for unadjusted' $\mathcal{T}_{\mathrm{L}}$ (FDA, 2023).

There is a line of research weakening Assumption 1 to censoring at random (Robins & Finkelstein, 2000; Lu & Tsiatis, 2011; Díaz et al., 2019), under which, however, the log-rank test is not valid and needs to be replaced by a weighted log-rank test that requires a correctly specified censoring distribution as the weights are inverse probabilities of censoring. Thus, the conditions and properties of weighted log-rank tests are not comparable with those of the log-rank test. Furthermore, the validity of weighted log-rank tests has only been established under simple randomization. The study of weighted log-rank tests under covariate-adaptive randomization is left for future work.

3. Covariate-adjusted log-rank test

Let $X \subset W$ be the vector containing observed baseline covariates to be adjusted in the construction of tests, with a nonsingular covariance matrix $\Sigma_X = \mathrm{var}(X)$. In this section, we develop a nonparametric covariate-adjusted log-rank test that has a simple and explicit formula, enjoys guaranteed efficiency gain over the log-rank test and is universally valid under all covariate-adaptive randomization schemes satisfying Condition 1.

To develop our covariate adjustment method, we first consider the following linearization of $\hat{U}_{\mathrm{L}}$ in (1):

$$\hat{U}_{\mathrm{L}} = U_{\mathrm{lin}} + n^{-1/2}o_p(1)\,U_{\mathrm{lin}} = \frac{1}{n}\sum_{i=1}^{n}\{I_i O_{i1} - (1-I_i)O_{i0}\},$$

$$O_{ij} = \int_0^{\tau}\{1-\mu(t)\}^j\{\mu(t)\}^{1-j}\{dN_{ij}(t) - Y_{ij}(t)p(t)\,\mathrm{d}t\}, \qquad j = 0, 1,$$

(Lin & Wei, 1989; Ye & Shao, 2020). Here $\mu(t) = E\{I_i \mid Y_i(t) = 1\}$, $p(t)\mathrm{d}t = E\{\mathrm{d}N_i(t)\}/E\{Y_i(t)\}$ and $o_p(1)$ denotes a term converging to 0 in probability as $n \to \infty$. Note that U_{lin} is an average of random variables that are independent and identically distributed under simple randomization. If we treat the O_{ij} in U_{lin} as outcomes and apply the generalized regression adjustment or augmentation (Cassel et al., 1976; Tsiatis et al., 2008) then we obtain the covariate-adjusted statistic

$$\begin{aligned} U_{\mathrm{Clin}} &= \frac{1}{n}\sum_{i=1}^{n}[I_i\{O_{i1} - (X_i - \bar{X})^{\mathrm{T}}\beta_1\} - (1 - I_i)\{O_{i0} - (X_i - \bar{X})^{\mathrm{T}}\beta_0\}] \\ &= U_{\mathrm{lin}} - \frac{1}{n}\sum_{i=1}^{n}\{I_i(X_i - \bar{X})^{\mathrm{T}}\beta_1 - (1 - I_i)(X_i - \bar{X})^{\mathrm{T}}\beta_0\}, \end{aligned} \tag{2}$$

where $\bar{X}$ is the sample mean of all the X_i, a^{T} is the transpose of vector a and $\beta_j = \sum_X^{-1}\mathrm{cov}(X_i, O_{ij})$ for $j = 0, 1$. Because the distribution of baseline covariate X_i is not affected by treatment, the last term on the right-hand side of (2) has mean 0. Under simple randomization, it follows from the theory of generalized regression (Cassel et al., 1976) that $\mathrm{var}(U_{\mathrm{Clin}}) \leqslant \mathrm{var}(U_{\mathrm{lin}})$, and thus the covariate adjusted U_{Clin} in (2) has a guaranteed efficiency gain over the unadjusted U_{lin}. This also holds under covariate-adaptive randomization; see Theorem S1 in the Supplementary Material.

To derive our covariate-adjusted procedure, it remains to find appropriate statistics to replace the O_{ij} and β_j in (2) because they involve unknown quantities. We consider the following sample analog of O_{ij}:

$$\hat{O}_{ij} = \int_0^{\tau}\frac{\bar{Y}_{1-j}(t)}{\bar{Y}(t)}\left\{\mathrm{d}N_{ij}(t) - Y_{ij}(t)\frac{\mathrm{d}\bar{N}(t)}{\bar{Y}(t)}\right\}, \qquad j = 0, 1, \tag{3}$$

with $\bar{N}(t) = \sum_{i=1}^{n} N_i(t)/n$. Using a correct form of $\hat{O}_{ij}$ is important, as it captures the true correlation between O_{ij} and X_i; see the discussion after Theorem S1 in the Supplementary Material. Replacing O_{ij} in (2) by the derived outcome $\hat{O}_{ij}$ in (3), we obtain the following covariate-adjusted version of $\hat{U}_{\mathrm{L}}$:

$$\begin{aligned} \hat{U}_{\mathrm{CL}} &= \frac{1}{n}\sum_{i=1}^{n}[I_i\{\hat{O}_{i1} - (X_i - \bar{X})^{\mathrm{T}}\hat{\beta}_1\} - (1 - I_i)\{\hat{O}_{i0} - (X_i - \bar{X})^{\mathrm{T}}\hat{\beta}_0\}] \\ &= \hat{U}_L - \frac{1}{n}\sum_{i=1}^{n}\{I_i(X_i - \bar{X})^{\mathrm{T}}\hat{\beta}_1 - (1 - I_i)(X_i - \bar{X})^{\mathrm{T}}\hat{\beta}_0\}. \end{aligned} \tag{4}$$

The second equality follows from the algebraic identity $\hat{U}_{\mathrm{L}} = n^{-1}\sum_{i=1}^{n}\{I_i\hat{O}_{i1} - (1 - I_i)\hat{O}_{i0}\}$,

$$\hat{\beta}_j = \left\{\sum_{i:I_i=j}(X_i - \bar{X}_j)(X_i - \bar{X}_j)^{\mathrm{T}}\right\}^{-1}\sum_{i:I_i=j}(X_i - \bar{X}_j)\hat{O}_{ij} \tag{5}$$

is a sample analog of $\beta_j = \sum_X^{-1}\mathrm{cov}(X_i, O_{ij})$ and $\bar{X}_j$ is the sample mean of the X_i with $I_i = j$. By Lemma S1 in the Supplementary Material, $\hat{\beta}_j$ in (5) converges to β_j in probability, which guarantees that $\hat{U}_{\rm CL}$ in (4) reduces the variability of $\hat{U}_{\rm L}$ in (1). Thus, we propose the covariate-adjusted log-rank test

$$\mathcal{T}_{\rm CL} = n^{1/2}\hat{U}_{\rm CL}/\hat{\sigma}_{\rm CL}, \tag{6}$$

where $\hat{\sigma}^2_{\rm CL} = \hat{\sigma}^2_{\rm L} - \pi(1-\pi)(\hat{\beta}_1+\hat{\beta}_0)^{\rm T}\hat{\Sigma}_X(\hat{\beta}_1+\hat{\beta}_0)$, whose form is suggested by $\sigma^2_{\rm CL}$ in Theorem 1, $\hat{\sigma}^2_{\rm L}$ is defined in (1) and $\hat{\Sigma}_X$ is the sample covariance matrix of all the X_i.

Asymptotic properties of covariate-adjusted log-rank test $\mathcal{T}_{\rm CL}$ in (6) are established in the following theorem. All technical proofs are given in the Supplementary Material. In what follows, $\xrightarrow{\rm D}$ and $\xrightarrow{\rm P}$, respectively, denote convergence in distribution and in probability, as $n \to \infty$.

Theorem 1. *Suppose that Condition 1 and Assumption 1 hold, and that all levels of Z_i used in covariate-adaptive randomization are included in X_i as a subvector. Then, the following results hold regardless of which covariate-adaptive randomization scheme is applied.*

(a) *Under the null H_0 or alternative hypothesis,*

$$n^{1/2}\{\hat{U}_{\rm CL} - (n_1\theta_1 - n_0\theta_0)/n\} \xrightarrow{\rm D} N(0, \sigma^2_{\rm CL}),$$

where $\theta_j = E(O_{ij})$, n_j is the number of patients in treatment j, $\sigma^2_{\rm CL} = \sigma^2_{\rm L} - \pi(1-\pi)(\beta_1 + \beta_0)^{\rm T}\Sigma_X(\beta_1+\beta_0)$ and $\sigma^2_{\rm L} = \pi{\rm var}(O_{i1}) + (1-\pi){\rm var}(O_{i0})$.

(b) *Under the null hypothesis H_0,*

$$\theta_1 = \theta_0 = 0, \qquad \hat{\sigma}^2_{\rm CL} \xrightarrow{\rm P} \sigma^2_{\rm CL} \quad \textit{and} \quad \mathcal{T}_{\rm CL} \xrightarrow{\rm D} N(0,1),$$

i.e., $\mathcal{T}_{\rm CL}$ is valid.

(c) *Under the local alternative hypothesis that $\theta_j = c_j n^{-1/2}$ with the c_j not depending on n and that $\lambda_1(t)/\lambda_0(t)$ is bounded and tends to 1 for every t,*

$$\mathcal{T}_{\rm CL} \xrightarrow{\rm D} N[\{\pi c_1 - (1-\pi)c_0\}/\sigma_{\rm CL}, 1].$$

The results under an alternative hypothesis in Theorem 1 are obtained without any specific model on the distribution of T_j or C_j, different from many published research articles that assume a specific model under an alternative hypothesis, such as the Cox proportional hazards model for T_j.

Theorem 1 shows that $\mathcal{T}_{\rm CL}$ in (6) is applicable to all randomization schemes satisfying Condition 1 with a universal formula, if all levels of Z_i are included in X_i. Tests with universal applicability are desirable for application, as the complication of using tailored formulas for different randomization schemes is avoided.

To show that $\mathcal{T}_{\rm CL}$ in (6) has a guaranteed efficiency gain over the benchmark $\mathcal{T}_{\rm L}$ in (1), we establish an asymptotic result for $\mathcal{T}_{\rm L}$ under covariate-adaptive randomization satisfying an additional condition.

Condition $1^\dagger$. As $n \to \infty$, $n^{1/2}(n_{z1}/n_z - \pi,\ z \in \mathcal{Z})^{\rm T} \mid Z_1, \ldots, Z_n \xrightarrow{\rm D} N(0, \Omega)$, where $\mathcal{Z}$ is the set containing all levels of Z, Ω is the diagonal matrix whose diagonal entries are $\nu/{\rm pr}(Z=z)$, $z \in \mathcal{Z}$, and $\nu \leqslant \pi(1-\pi)$ is a known constant depending on the randomization scheme.

THEOREM 2. *Suppose that Conditions 1 and $1^\dagger$ and Assumption 1 hold. Then the following results hold.*

(a) *Under the null H_0 or alternative hypothesis,*

$$n^{1/2}\{\hat{U}_{\mathrm{L}} - (n_1\theta_1 - n_0\theta_0)/n\} \xrightarrow{\mathrm{D}} N\{0, \sigma^2_{\mathrm{L}}(\nu)\},$$

where n_j and θ_j are given in Theorem 1, $\sigma^2_{\mathrm{L}}(\nu) = \sigma^2_{\mathrm{L}} - \{\pi(1-\pi) - \nu\}\mathrm{var}\{E(O_{i1} \mid Z_i) + E(O_{i0} \mid Z_i)\}$ for ν given in Condition $1^\dagger$ and σ^2_{L} is defined in Theorem 1.

(b) *Under the null hypothesis H_0,*

$$\theta_1 = \theta_0 = 0, \qquad \hat{\sigma}^2_{\mathrm{L}} \xrightarrow{\mathrm{P}} \sigma^2_{\mathrm{L}} \quad \text{and} \quad \mathcal{T}_{\mathrm{L}} \xrightarrow{\mathrm{D}} N\{0, \sigma^2_{\mathrm{L}}(\nu)/\sigma^2_{\mathrm{L}}\}.$$

Hence, $\mathcal{T}_{\mathrm{L}}$ is conservative unless $\nu = \pi(1-\pi)$ or $E(O_{i1} \mid Z_i) + E(O_{i0} \mid Z_i) = 0$ almost surely under H_0.

(c) *Under the local alternative hypothesis in Theorem 1(c),*

$$\mathcal{T}_{\mathrm{L}} \xrightarrow{\mathrm{D}} N[\{\pi c_1 - (1-\pi)c_0\}/\sigma_{\mathrm{L}}, \sigma^2_{\mathrm{L}}(\nu)/\sigma^2_{\mathrm{L}}].$$

Under simple randomization, Condition $1^\dagger$ holds with $\nu = \pi(1-\pi)$ and, hence, Theorem 2 also applies with $\sigma^2_{\mathrm{L}}(\nu) = \sigma^2_{\mathrm{L}}$. Under the local alternative specified in Theorem 1(c) with $\pi c_1 - (1-\pi)c_0 \neq 0$, by Theorems 1(c) and 2(c), Pitman's asymptotic relative efficiency of $\mathcal{T}_{\mathrm{CL}}$ in (6) with respect to the benchmark $\mathcal{T}_{\mathrm{L}}$ in (1) is $\sigma^2_{\mathrm{L}}/\sigma^2_{\mathrm{CL}} = 1 + \pi(1-\pi)(\beta_1 + \beta_0)^{\mathrm{T}}\Sigma_X(\beta_1+\beta_0)/\sigma^2_{\mathrm{CL}} \geqslant 1$ with the strict inequality holding unless $\beta_1 + \beta_0 = 0$. Thus, $\mathcal{T}_{\mathrm{CL}}$ has a guaranteed efficiency gain over $\mathcal{T}_{\mathrm{L}}$ under simple randomization.

Under covariate-adaptive randomization satisfying Condition $1^\dagger$ with $\nu < \pi(1-\pi)$, Theorem 2(b) shows that $\mathcal{T}_{\mathrm{L}}$ is not valid, but conservative, as $\sigma^2_{\mathrm{L}}(\nu) < \sigma^2_{\mathrm{L}}$ unless $E(O_{i1} \mid Z_i) + E(O_{i0} \mid Z_i) = 0$ almost surely under H_0, which holds under some extreme scenarios, e.g., Z used for randomization is independent of the outcome. This conservativeness can be corrected by a multiplication factor $\hat{r}(\nu) \xrightarrow{\mathrm{P}} \sigma_{\mathrm{L}}/\sigma_{\mathrm{L}}(\nu)$ under H_0. The resulting $\hat{r}(\nu)\mathcal{T}_{\mathrm{L}}$ is the modified log-rank test in Ye & Shao (2020), which is valid and always more powerful than $\mathcal{T}_{\mathrm{L}}$. Under the local alternative specified in Theorem 1(c) with $\pi c_1 - (1-\pi)c_0 \neq 0$, Pitman's asymptotic relative efficiency of $\hat{r}(\nu)\mathcal{T}_{\mathrm{L}}$ with respect to $\mathcal{T}_{\mathrm{CL}}$ in (6) is

$$\begin{aligned}\frac{\sigma^2_{\mathrm{L}}(\nu)}{\sigma^2_{\mathrm{CL}}} &= 1 + (\beta_1+\beta_0)^{\mathrm{T}}[\pi(1-\pi)E\{\mathrm{var}(X_i \mid Z_i)\} + \nu\mathrm{var}\{E(X_i \mid Z_i)\}]\frac{\beta_1+\beta_0}{\sigma^2_{\mathrm{CL}}}\\ &\geqslant 1\end{aligned}$$

with the strict inequality holding unless $\beta_1 + \beta_0 = 0$, e.g., X_i is uncorrelated with O_{ij}, or $\nu = 0$ and $E\{\mathrm{cov}(X_i, O_{i1} \mid Z_i) + \mathrm{cov}(X_i, O_{i0} \mid Z_i)\} = 0$, e.g., covariates in X_i, but not in Z_i, are uncorrelated with O_{ij} conditioned on Z_i. Hence, the adjusted $\mathcal{T}_{\mathrm{CL}}$ has a guaranteed efficiency gain over both the log-rank test $\mathcal{T}_{\mathrm{L}}$ and modified log-rank test $\hat{r}(\nu)\mathcal{T}_{\mathrm{L}}$ under any covariate-adaptive randomization schemes satisfying Conditions 1 and $1^\dagger$.

The Pocock–Simon minimization satisfies Condition 1, but not necessarily Condition $1^\dagger$ as the I_i are correlated across strata. Hence, under the Pocock–Simon minimization, Theorem 2 is not applicable and $\mathcal{T}_{\mathrm{L}}$ may not be valid, whereas $\mathcal{T}_{\mathrm{CL}}$ is valid according to Theorem 1, another advantage of covariate adjustment.

In the numerator of (6) $\hat{U}_{\rm CL}$ is the same as the augmented score in Lu & Tsiatis (2008), which shares the same idea as those in Tsiatis et al. (2008) and Zhang et al. (2008) for non-censored data. However, the denominator $\hat{\sigma}_{\rm CL}$ in (6) is different from that used by Lu & Tsiatis (2008). The key difference between our result on guaranteed efficiency gain and the result in Lu & Tsiatis (2008) is that our result is obtained under covariate-adaptive randomization and an alternative hypothesis without any specific model on the distribution of T_j or C_j, whereas the result in Lu & Tsiatis (2008) is for simple randomization and an alternative hypothesis under a correctly specified Cox proportional hazards model for T_j.

After testing H_0, it is often of interest to estimate and construct a confidence interval for an effect size (Lu & Tsiatis, 2008; Parast et al., 2014; Zhang, 2015; Díaz et al., 2019). A commonly considered effect size is the hazard ratio e^{θ} under the Cox proportional hazards model $\lambda_1(t) = \lambda_0(t)e^{\theta}$. The hazard ratio e^{θ} is interpretable only when the Cox proportional hazards model is correctly specified. Thus, in the rest of this section we consider covariate-adjusted estimation and a confidence interval for θ, assuming that $\lambda_1(t) = \lambda_0(t)e^{\theta}$.

Without using any covariate, the score from the partial likelihood under model $\lambda_1(t) = \lambda_0(t)e^{\theta}$ is

$$\hat{U}_{\rm L}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{I_i - \frac{e^{\vartheta}\,\bar{Y}_1(t)}{e^{\vartheta}\,\bar{Y}_1(t) + \bar{Y}_0(t)}\right\} {\rm d}N_i(t).$$

The maximum partial likelihood estimator $\hat{\theta}_{\rm L}$ of θ is a solution to $\hat{U}_{\rm L}(\vartheta) = 0$. Using the idea in (4) with X_i containing all levels of Z_i used in covariate-adaptive randomization, our covariate-adjusted score is

$$\hat{U}_{\rm CL}(\vartheta) = \hat{U}_{\rm L}(\vartheta) - \frac{1}{n}\sum_{i=1}^{n}\{I_i(X_i - \bar{X})^{\rm T}\hat{\beta}_1(\hat{\theta}_{\rm L}) - (1 - I_i)(X_i - \bar{X})^{\rm T}\hat{\beta}_0(\hat{\theta}_{\rm L})\},$$

where, for $j = 0, 1$, $\hat{\beta}_j(\vartheta)$ is equal to $\hat{\beta}_j$ in (5) with $\hat{O}_{ij}$ replaced by

$$\hat{O}_{ij}(\vartheta) = \int_0^{\tau}\frac{\{e^{\vartheta}\,\bar{Y}_1(t)\}^{(1-j)}\{\bar{Y}_0(t)\}^{j}}{e^{\vartheta}\,\bar{Y}_1(t) + \bar{Y}_0(t)}\left\{dN_{ij}(t) - \frac{Y_{ij}(t)e^{j\vartheta}\,d\bar{N}(t)}{e^{\vartheta}\,\bar{Y}_1(t) + \bar{Y}_0(t)}\right\}.$$

Solving $\hat{U}_{\rm CL}(\vartheta) = 0$ gives the covariate-adjusted estimator $\hat{\theta}_{\rm CL}$. As $\hat{U}_{\rm CL}(\theta)$ has reduced variability compared to $\hat{U}_{\rm L}(\theta)$, and $\partial\hat{U}_{\rm CL}(\vartheta)/\partial\vartheta = \partial\hat{U}_{\rm L}(\vartheta)/\partial\vartheta$, by a standard argument for M-estimators, $\hat{\theta}_{\rm CL}$ is guaranteed to have smaller variance than $\hat{\theta}_{\rm L}$. It is established in the Supplementary Material that $n^{1/2}(\hat{\theta}_{\rm CL} - \theta) \xrightarrow{\rm D} N\{0, \sigma^2(\theta)\}$ under any covariate-adaptive randomization satisfying Condition 1, with $\sigma^2(\theta)$ given in Theorem S2 in the Supplementary Material. An asymptotic confidence interval for θ can be obtained based on this result and a consistent estimator of $\sigma^2(\theta)$ given by

$$[g(\hat{\theta}_{\rm CL}) - \pi(1-\pi)\{\hat{\beta}_1(\hat{\theta}_{\rm L}) + \hat{\beta}_0(\hat{\theta}_{\rm L})\}^{\rm T}\hat{\Sigma}_X\{\hat{\beta}_1(\hat{\theta}_{\rm L}) + \hat{\beta}_0(\hat{\theta}_{\rm L})\}]/\{g(\hat{\theta}_{\rm CL})\}^2,$$

where $g(\vartheta) = -\partial\hat{U}_{\rm L}(\vartheta)/\partial\vartheta$.

4. Covariate-adjusted stratified log-rank test

The stratified log-rank test (Peto et al., 1976) is a weighted average of the stratum-specific log-rank test statistics with finitely many strata constructed using a discrete baseline covariate. We consider stratification with all levels of Z_i. Results can be obtained similarly for stratifying on more levels than those of Z_i or fewer levels than those of Z_i with levels of Z_i not used in stratification included in X_i. Here, we remove the part of X_i that can be linearly represented by Z_i and still denote the remaining as X_i. As such, it is reasonable to assume that $E\{\mathrm{var}(X_i \mid Z_i)\}$ is positive definite.

The stratified log-rank test using levels of Z_i as strata is

$$\mathcal{T}_{\mathrm{SL}} = n^{1/2}\hat{U}_{\mathrm{SL}}/\hat{\sigma}_{\mathrm{SL}}, \tag{7}$$

where

$$\hat{U}_{\mathrm{SL}} = \frac{1}{n}\sum_{z}\sum_{i:Z_i=z}\int_0^{\tau}\left\{I_i - \frac{\bar{Y}_{z1}(t)}{\bar{Y}_z(t)}\right\}\mathrm{d}N_i(t), \hat{\sigma}^2_{\mathrm{SL}} \quad = \frac{1}{n}\sum_{z}\sum_{i:Z_i=z}\int_0^{\tau}\frac{\bar{Y}_{z1}(t)\bar{Y}_{z0}(t)}{\bar{Y}_z(t)^2}\,\mathrm{d}N_i(t),$$

$\bar{Y}_{z1}(t) = \sum_{i:Z_i=z} I_i Y_i(t)/n$, $\bar{Y}_{z0}(t) = \sum_{i:Z_i=z}(1-I_i)Y_i(t)/n$ and $\bar{Y}_z(t) = \bar{Y}_{z1}(t) + \bar{Y}_{z0}(t)$.

With stratification, $\mathcal{T}_{\mathrm{SL}}$ in (7) actually tests the null hypothesis $\widetilde{H}_0\colon \lambda_1(t \mid z) = \lambda_0(t \mid z)$ for all (t, z), where $\lambda_j(t \mid z)$ is the hazard function of T_j conditional on $Z = z$. Hypothesis $\widetilde{H}_0$ may be stronger than $H_0\colon \lambda_1(t) = \lambda_0(t)$ for all t, the null hypothesis for unstratified log-rank test $\mathcal{T}_{\mathrm{L}}$ and its adjustment $\mathcal{T}_{\mathrm{CL}}$ considered in § 2–§ 3. In some scenarios, $\widetilde{H}_0 = H_0$. For example, the two hypotheses are the same when there exists a transformation model $h\{\mathrm{pr}(T_0 \geqslant t \mid W)\} = \theta + h\{\mathrm{pr}(T_1 \geqslant t \mid W)\}$ for all (t, W) and an unknown constant θ, where h is an increasing function that is possibly unknown (Cheng et al., 1995). This transformation model includes many commonly used semiparametric models as special cases, for example the Cox proportional hazards model with $h(s) = -\log\{-\log(s)\}$.

To further adjust for baseline covariate X_i, we still linearize $\hat{U}_{\mathrm{SL}}$ as (Ye & Shao, 2020)

$$\hat{U}_{\mathrm{SL}} = \frac{1}{n}\sum_{z}\sum_{i:Z_i=z}\{I_i O_{zi1} - (1-I_i)O_{zi0}\} + n^{-1/2}o_p(1),$$

where

$$O_{zij} = \int_0^{\tau}\{1-\mu_z(t)\}^j\{\mu_z(t)\}^{1-j}\{\mathrm{d}N_{ij}(t) - Y_{ij}(t)p_z(t)\,\mathrm{d}t\}, \qquad j = 0, 1,$$

with

$$p_z(t)\mathrm{d}t = \frac{E\{\mathrm{d}N_i(t) \mid Z_i = z\}}{E\{Y_i(t) \mid Z_i = z\}} \quad \text{and} \quad \mu_z(t) = E(I_i \mid Y_i(t) = 1, Z_i = z).$$

Following the same idea as in § 3, we apply the generalized regression adjustment by using

$$\hat{O}_{zij} = \int_0^{\tau}\frac{\bar{Y}_{z(1-j)}(t)}{\bar{Y}_z(t)}\left\{\mathrm{d}N_{ij}(t) - Y_{ij}(t)\frac{\mathrm{d}\bar{N}_z(t)}{\bar{Y}_z(t)}\right\}, \qquad j = 0, 1,$$

as derived outcomes, where $\bar{N}_z(t) = \sum_{i:Z_i=z} N_i(t)/n$. The resulting covariate-adjusted version of $\hat{U}_{\mathrm{SL}}$ is

$$\begin{aligned}\hat{U}_{\mathrm{CSL}} &= \frac{1}{n}\sum_z \sum_{i:Z_i=z} [I_i\{\hat{O}_{zi1} - (X_i - \bar{X}_z)^{\mathrm{T}}\hat{\gamma}_1\} - (1-I_i)\{\hat{O}_{zi0} - (X_i - \bar{X}_z)^{\mathrm{T}}\hat{\gamma}_0\}] \\ &= \hat{U}_{\mathrm{SL}} - \frac{1}{n}\sum_z \sum_{i:Z_i=z} \{I_i(X_i - \bar{X}_z)^{\mathrm{T}}\hat{\gamma}_1 - (1-I_i)(X_i - \bar{X}_z)^{\mathrm{T}}\hat{\gamma}_0\},\end{aligned}$$

where the second equality follows from the algebraic identity $\hat{U}_{\mathrm{SL}} = n^{-1}\sum_z \sum_{i:Z_i=z}\{I_i\hat{O}_{zi1} - (1-I_i)\hat{O}_{zi0}\}$, $\bar{X}_z$ is the sample mean of the X_i with $Z_i = z$,

$$\hat{\gamma}_j = \left\{\sum_z \sum_{i:I_i=j,\,Z_i=z} (X_i - \bar{X}_{zj})(X_i - \bar{X}_{zj})^{\mathrm{T}}\right\}^{-1} \sum_z \sum_{i:I_i=j,\,Z_i=z} (X_i - \bar{X}_{zj})\hat{O}_{zij},$$

converging to a limit value γ_j in probability, and $\bar{X}_{zj}$ is the sample mean of the X_i with $Z_i = z$ and treatment j, $j = 0, 1$. Our proposed covariate-adjusted stratified log-rank test is

$$\mathcal{T}_{\mathrm{CSL}} = n^{1/2}\hat{U}_{\mathrm{CSL}}/\hat{\sigma}_{\mathrm{CSL}}, \tag{8}$$

where $\hat{\sigma}^2_{\mathrm{CSL}} = \hat{\sigma}^2_{\mathrm{SL}} - \pi(1-\pi)(\hat{\gamma}_1 + \hat{\gamma}_0)^{\mathrm{T}}\{\sum_z (n_z/n)\hat{\Sigma}_{X|z}\}(\hat{\gamma}_1 + \hat{\gamma}_0)$ and $\hat{\Sigma}_{X|z}$ is the sample covariance matrix of the X_i within stratum z.

The following theorem establishes the asymptotic properties of the stratified log-rank test $\mathcal{T}_{\mathrm{SL}}$ and covariate-adjusted stratified log-rank test $\mathcal{T}_{\mathrm{CSL}}$.

Theorem 3. *Suppose that Condition 1 holds and that $C_I \perp T_I \mid (I, Z)$. Then, the following results hold regardless of which covariate-adaptive randomization is applied.*

(a) *Under the null $\widetilde{H}_0$ or alternative hypothesis,*

$$n^{1/2}\left\{\hat{U}_{\mathrm{CSL}} - \sum_z (n_{z1}\theta_{z1} - n_{z0}\theta_{z0})/n\right\} \xrightarrow{\mathrm{D}} N(0, \sigma^2_{\mathrm{CSL}}),$$

and the same result holds with $\hat{U}_{\mathrm{CSL}}$ and σ^2_{CSL} replaced by $\hat{U}_{\mathrm{SL}}$ and σ^2_{SL}, respectively, where $\theta_{zj} = E(O_{zij} \mid Z_i = z)$, n_{zj} is the number of patients with treatment j in stratum z, $j = 0, 1$, $\sigma^2_{\mathrm{CSL}} = \sigma^2_{\mathrm{SL}} - \pi(1-\pi)(\gamma_1 + \gamma_0)^{\mathrm{T}}E\{\mathrm{var}(X_i \mid Z_i)\}(\gamma_1 + \gamma_0)$ and $\sigma^2_{\mathrm{SL}} = \sum_z \mathrm{pr}(Z_i = z)\{\pi\mathrm{var}(O_{zi1} \mid Z_i = z) + (1-\pi)\mathrm{var}(O_{zi0} \mid Z_i = z)\}$.

(b) *Under the null hypothesis $\widetilde{H}_0$,*

$$\theta_{z1} = \theta_{z0} = 0 \quad \textit{for any } z, \qquad \hat{\sigma}^2_{\mathrm{SL}} \xrightarrow{\mathrm{P}} \sigma^2_{\mathrm{SL}}, \qquad \hat{\sigma}^2_{\mathrm{CSL}} \xrightarrow{\mathrm{P}} \sigma^2_{\mathrm{CSL}},$$

$$\mathcal{T}_{\mathrm{SL}} \xrightarrow{\mathrm{D}} N(0,1) \quad \textit{and} \quad \mathcal{T}_{\mathrm{CSL}} \xrightarrow{\mathrm{D}} N(0,1),$$

i.e., both $\mathcal{T}_{\mathrm{SL}}$ and $\mathcal{T}_{\mathrm{CSL}}$ are valid for testing null hypothesis $\widetilde{H}_0$.

(c) *Under the local alternative hypothesis that $\theta_{zj} = c_{zj}n^{-1/2}$ with the c_{zj} not depending on n and that $\lambda_1(t \mid z)/\lambda_0(t \mid z)$ is bounded and tends to 1 for every t and z,*

$$\mathcal{T}_{\mathrm{CSL}} \xrightarrow{\mathrm{D}} N\left(\sum_z \mathrm{pr}(Z = z)\{\pi c_{z1} - (1-\pi)c_{z0}\}/\sigma_{\mathrm{CSL}}, 1\right),$$

and the same result holds with $\mathcal{T}_{\mathrm{CSL}}$ and σ_{CSL} replaced by $\mathcal{T}_{\mathrm{SL}}$ and σ_{SL}, respectively.

Like $\mathcal{T}_{\text{CL}}$ in (6), both $\mathcal{T}_{\text{SL}}$ in (7) and $\mathcal{T}_{\text{CSL}}$ in (8) are applicable to all covariate-adaptive randomization schemes with universal formulas, i.e., they achieve the universal applicability. In terms of Pitman's asymptotic efficiency under the local alternative specified in Theorem 3(c), $\mathcal{T}_{\text{CSL}}$ is always more efficient than $\mathcal{T}_{\text{SL}}$, since $\sigma^2_{\text{CSL}} \leqslant \sigma^2_{\text{SL}}$ with the strict inequality holding unless $\gamma_1 + \gamma_0 = 0$.

The condition $C_I \perp T_I \mid (I, Z)$ in Theorem 3 for the stratified log-rank test and its adjustment is in general not comparable with Theorem 1(c) for the unstratified log-rank test.

Is $\mathcal{T}_{\text{SL}}$ or $\mathcal{T}_{\text{CSL}}$ more efficient than the unstratified log-rank test $\mathcal{T}_{\text{L}}$? The answer is not clear because, firstly, the null hypotheses $\widetilde{H}_0$ and H_0 may be different, as we discussed earlier, and secondly, even if $\widetilde{H}_0 = H_0$, under the alternative, the asymptotic mean $(n_1\theta_1 - n_0\theta_0)/n$ of $\hat{U}_{\text{L}}$ may not be comparable with the asymptotic mean $\sum_z (n_{z1}\theta_{z1} - n_{z0}\theta_{z0})/n$ of $\hat{U}_{\text{SL}}$ or $\hat{U}_{\text{CSL}}$. In fact, the indefiniteness of relative efficiency between the stratified and unstratified log-rank tests is a standing problem in the literature.

There is also no definite answer when comparing the efficiencies of $\mathcal{T}_{\text{CL}}$ and the stratified $\mathcal{T}_{\text{CSL}}$.

Similar to the discussion at the end of § 3, after testing hypothesis $\widetilde{H}_0$, we can obtain a covariate-adjusted confidence interval for the effect size θ under a stratified Cox proportional hazards model $\lambda_{1z}(t) = \lambda_{0z}(t)e^{\theta}$ for every z; see the Supplementary Material for further details.

5. Simulations

To supplement the theory and examine finite sample Type-I error and power of tests $\mathcal{T}_{\text{L}}$, $\mathcal{T}_{\text{CL}}$, $\mathcal{T}_{\text{SL}}$ and $\mathcal{T}_{\text{CSL}}$, we carry out a simulation study under the following four cases/models.

Case I. The conditional hazard function follows a Cox model, $\lambda_j(t \mid W) = (\log 2)\exp(-\theta j + \eta^{\mathrm{T}} W)$ for $j = 0, 1$, where θ denotes a scalar parameter, $\eta = (0.5, 0.5, 0.5)^{\mathrm{T}}$ and W is a three-dimensional covariate vector following the three-dimensional standard normal distribution. The censoring variables C_0 and C_1 follow a uniform distribution on the interval $(10, 40)$ and are independent of W.

Case II. The conditional hazard function is the same as that in Case I. Conditional on W and treatment assignment j, $C_j - (3 - 3j)$ follows a standard exponential distribution.

Case III. We have $T_j = \exp(-\theta j + \eta^{\mathrm{T}} W) + \mathcal{E}$, $j = 0, 1$, where θ, η and W are the same as in Case I, and $\mathcal{E}$ is a random variable independent of (C_1, C_0, W) and has the standard exponential distribution. The setting for censoring is the same as that in Case I.

Case IV. The models for the T_j and C_j are the same as those in Cases III and II, respectively.

In this simulation, the significance level $\alpha = 5\%$, the target treatment assignment proportion $\pi = 0.5$, the overall sample size $n = 500$, the null hypothesis $H_0 : \theta = 0$, and $\widetilde{H}_0 = H_0$ since a transformation model described in § 4 holds in Cases I–IV. Three randomization schemes are considered: simple randomization, stratified permuted block randomization with block size 4 and levels of Z as strata, and the Pocock–Simon minimization assigning a patient with probability 0.8 to the preferred arm minimizing the sum of balance scores over marginal levels of Z, where Z is the two-dimensional vector whose first component is a two-level discretized first component of W and the second component is a three-level discretized second component of W. For stratified log-rank tests, levels of Z are used as strata.

Table 1. *Type-I errors (in percentages) based on 10 000 simulations*

Case	Randomization	$\mathcal{T}_{\mathrm{L}}$	$\mathcal{T}_{\mathrm{CL}}$	$\mathcal{T}_{\mathrm{SL}}$	$\mathcal{T}_{\mathrm{CSL}}$
I	Simple	4.91	5.16	4.86	4.78
	Permuted block	3.25	5.22	4.80	4.85
	Minimization	3.40	5.43	5.02	5.23
II	Simple	5.39	5.14	5.00	4.97
	Permuted block	3.59	5.03	4.94	4.82
	Minimization	4.01	5.23	5.11	5.28
III	Simple	5.07	5.43	5.27	5.16
	Permuted block	2.29	4.79	4.76	4.82
	Minimization	2.88	5.43	5.23	5.52
IV	Simple	5.41	5.30	5.39	5.21
	Permuted block	4.44	5.48	5.10	5.49
	Minimization	4.21	5.18	5.04	5.06

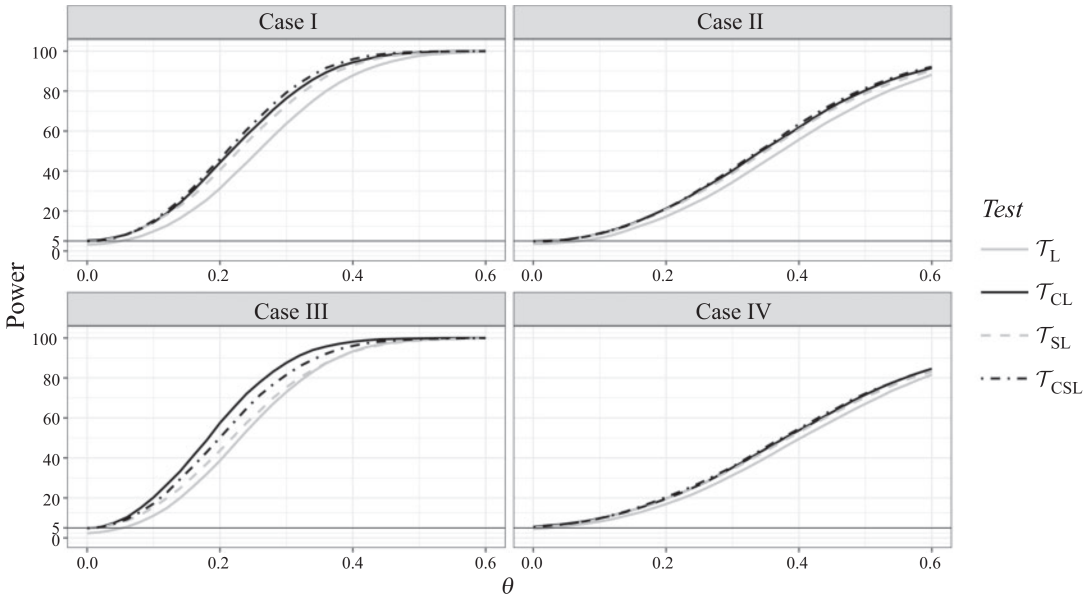

Fig. 1. Power curves based on 10 000 simulations.

For covariate adjustment, X is the vector containing Z and the third component of W for $\mathcal{T}_{\mathrm{CL}}$, and X is the third component of W for $\mathcal{T}_{\mathrm{CSL}}$.

Based on 10 000 simulations, Type-I error rates for four tests under four cases and three randomization schemes are shown in Table 1. The results agree with our theory. For $\mathcal{T}_{\mathrm{CL}}$, $\mathcal{T}_{\mathrm{SL}}$ and $\mathcal{T}_{\mathrm{CSL}}$, there is no substantial difference among the three randomization schemes. The log-rank test $\mathcal{T}_{\mathrm{L}}$ preserves the 5% rate under simple randomization, but it is conservative under stratified permuted block randomization and minimization.

Based on 10 000 simulations, power curves of four tests for θ ranging from 0 to 0.6, under four cases and stratified permuted block randomization are plotted in Fig. 1. Similar figures for simple randomization and minimization are given in the Supplementary Material. In all cases, the power curves of covariate-adjusted tests $\mathcal{T}_{\mathrm{CL}}$ and $\mathcal{T}_{\mathrm{CSL}}$ are better than those of unadjusted tests $\mathcal{T}_{\mathrm{L}}$ and $\mathcal{T}_{\mathrm{SL}}$, especially the benchmark $\mathcal{T}_{\mathrm{L}}$. Under Cox's model, $\mathcal{T}_{\mathrm{CSL}}$ is better than $\mathcal{T}_{\mathrm{CL}}$, but not necessarily under the non-Cox model. The stratified $\mathcal{T}_{\mathrm{SL}}$ is mostly better than the unstratified $\mathcal{T}_{\mathrm{L}}$, but unlike $\mathcal{T}_{\mathrm{CL}}$ and $\mathcal{T}_{\mathrm{CSL}}$, there is no guaranteed efficiency gain, e.g., case III when $\theta > 0.4$. The difference in censoring model also has some effect.

More simulation results can be found in the Supplementary Material.

6. A real data application

We apply four tests $\mathcal{T}_{\mathrm{L}}$, $\mathcal{T}_{\mathrm{CL}}$, $\mathcal{T}_{\mathrm{SL}}$ and $\mathcal{T}_{\mathrm{CSL}}$ to the data from the AIDS Clinical Trials Group Study 175, ACTG 175, a randomized controlled trial evaluating antiretroviral treatments in adults infected with human immunodeficiency virus type 1 whose CD4 cell counts were from 200 to 500 per cubic millimeter (Hammer et al., 1996). The primary endpoint was time to a composite event defined as a $\geqslant 50\%$ decline in the CD4 cell count, an AIDS-defining event, or death. Stratified permuted block randomization with equal allocation was applied with covariate Z having three levels related with the length of prior antiretroviral therapy: $Z = 1$, 2 and 3, representing 0 weeks, between 1 to 52 weeks and more than 52 weeks of prior antiretroviral therapy, respectively. The dataset is publicly available in the `R` package `speff2trial` (R Development Core Team, 2024).

We focus on the comparison of treatment 0 (zidovudine) versus treatment 1 (didanosine). For stratified log-rank test $\mathcal{T}_{\mathrm{SL}}$, the three-level Z is used as the stratification variable. For covariate adjustment, two additional prognostic baseline covariates are considered as X: the baseline CD4 cell count and the number of days receiving antiretroviral therapy prior to treatment. In addition to testing treatment effect for all patients, a subgroup analysis with Z strata as subgroups is also of interest because responses to antiretroviral therapy may vary according to the extent of prior drug exposure. Within each subgroup defined by Z, the stratified tests become the same as their unstratified counterparts, and thus we only apply tests $\mathcal{T}_{\mathrm{L}}$ and $\mathcal{T}_{\mathrm{CL}}$ in the subgroup analysis.

Table 2 reports the number of patients, numerator and denominator of each test, and a p-value for testing with all patients or with a subgroup. The effect of covariate adjustment is clear: for the covariate-adjusted tests, the standard errors $\hat{\sigma}_{\mathrm{CL}}$ and $\hat{\sigma}_{\mathrm{CSL}}$ are smaller than $\hat{\sigma}_{\mathrm{L}}$ and $\hat{\sigma}_{\mathrm{SL}}$ in all analyses.

For the analysis based on all patients, all four tests significantly reject the null hypothesis H_0 of the no-treatment effect. In the subgroup analysis, the p-values are adjusted using Bonferroni's correction to control for the familywise error rate. From Table 2, p-values in the subgroup analysis are substantially larger than those in the analysis of all patients, because of reduced sample sizes as well as Bonferroni's correction. The empirical result in this example illustrates the benefit of covariate adjustment in testing when the sample size is not very large. Using the adjusted log-rank test $\mathcal{T}_{\mathrm{CL}}$, together with the estimated effect size and its standard error shown in Table 2, we can conclude the superiority of treatment 1 for both $Z = 1$ and $Z = 3$, which is consistent with the evidence of Hammer et al. (1996).

Acknowledgement

We would like to thank all reviewers for useful comments and suggestions. Our research was supported by the National Natural Science Foundation of China and the U.S. National Science Foundation. Shao is also affiliated with the East China Normal University.

Supplementary material

The Supplementary Material contains all technical proofs and some additional results.

Table 2. *Statistics for the ACTG 175 example*

			Subgroup	
	All patients	$Z = 1$	$Z = 2$	$Z = 3$
Number of patients	1093	461	198	434
Log-rank test				
$n^{1/2}\hat{U}_{\mathrm{L}}$	-1.223	-0.542	-0.144	-1.292
$\hat{\sigma}_{\mathrm{L}}$	0.265	0.235	0.270	0.290
p-value (adjusted for subgroup analysis)	< 0.001	0.064	1	< 0.001
Estimated θ	-0.528	-0.455	-0.140	-0.740
Standard error of the estimated θ	0.116	0.199	0.263	0.171
Covariate-adjusted log-rank test				
$n^{1/2}\hat{U}_{\mathrm{CL}}$	-1.273	-0.553	-0.129	-1.382
$\hat{\sigma}_{\mathrm{CL}}$	0.257	0.230	0.265	0.282
p-value (adjusted for subgroup analysis)	< 0.001	0.049	1	< 0.001
Estimated θ	-0.550	-0.464	-0.127	-0.793
Standard error of the estimated θ	0.113	0.195	0.257	0.166
Stratified log-rank test				
$n^{1/2}\hat{U}_{\mathrm{SL}}$	-1.228			
$\hat{\sigma}_{\mathrm{SL}}$	0.264			
p-value	< 0.001			
Estimated θ	-0.531			
Standard error of the estimated θ	0.116			
Covariate-adjusted stratified log-rank test				
$n^{1/2}\hat{U}_{\mathrm{CSL}}$	-1.284			
$\hat{\sigma}_{\mathrm{CSL}}$	0.258			
p-value	< 0.001			
Estimated θ	-0.556			
Standard error of the estimated θ	0.113			

Here θ denotes the log hazard ratio for all patients and for each subgroup.

References

Baldi Antognini, A. & Zagoraiou, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli* **21**, 881–908.

Cassel, C. M., Särndal, C. E. & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–20.

Cheng, S., Wei, L. & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–45.

Ciolino, J. D., Palac, H. L., Yang, A., Vaca, M. & Belli, H. M. (2019). Ideal vs. real: a systematic review on handling covariates in randomized controlled trials. *BMC Med. Res. Methodol.* **19**, 136.

Díaz, I., Colantuoni, E., Hanley, D. F. & Rosenblum, M. (2019). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Anal.* **25**, 439–68.

DiRienzo, A. G. & Lagakos, S. W. (2002). Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *J. R. Statist. Soc.* B **63**, 745–57.

EMA (2015). Guideline on adjustment for baseline covariates in clinical trials, EMA/CHMP/295050/2013. London, UK: European Medicines Agency (EMA). https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf. Accessed August 3, 2023.

FDA (2023). Adjusting for covariates in randomized clinical trials for drugs and biological products. Guidance for Industry. Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, Food and Drug Administration (FDA), U.S. Department of Health and Human Services. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products. Accessed August 3 2023.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M. et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New Engl. J. Med.* **335**, 1081–90.

ICH E9. (1998). Statistical principles for clinical trials E9. International Council for Harmonisation (ICH). https://database.ich.org/sites/default/files/E9_Guideline.pdf. Accessed August 3, 2023.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. New York: John Wiley,.

Kong, F. H. & Slud, E. (1997). Robust covariate-adjusted logrank tests. *Biometrika* **84**, 847–62.

Lin, D. Y. & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074–8.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Statist.* **7**, 295–318.

Lu, X. & Tsiatis, A. A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* **95**, 679–94.

Lu, X. & Tsiatis, A. A. (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. *Lifetime Data Anal.* **17**, 566–93.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–70.

Moore, K. L. & van der Laan, M. J. (2009). Increasing power in randomized trials with right censored outcomes through covariate adjustment. *J. Biopharm. Statist.* **19**, 1099–131.

Parast, L., Tian, L. & Cai, T. (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *J. Am. Statist. Assoc.* **109**, 384–94.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. & Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* **34**, 585–612.

Pocock, S. J. & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31**, 103–15.

R Development Core Team (2024). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Robins, J. M. & Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56**, 779–88.

Schulz, K. F. & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* **359**, 515–19.

Shao, J. (2021). Inference for covariate-adaptive randomization: aspects of methodology and theory (with discussions). *Statist. Theory Rel. Fields* **5**, 172–86.

Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* **15**, 443–53.

Taves, D. R. (2010). The use of minimization in clinical trials. *Contemp. Clin. Trials* **31**, 180–4.

Tsiatis, A. A., Davidian, M., Zhang, M. & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statist. Med.* **27**, 4658–77.

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M. & Rosenblum, M. (2023). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *J. Am. Statist. Assoc.* **118**, 1152–63.

Ye, T. & Shao, J. (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *J. R. Statist. Soc.* B **82**, 1301–23.

Ye, T., Shao, J., Yi, Y. & Zhao, Q. (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *J. Am. Statist. Assoc.*, doi: 10.1080/01621459.2022.2049278.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *J. Chronic Dis.* **27**, 365–75.

Zhang, M. (2015). Robust methods to improve efficiency and reduce bias in estimating survival curves in randomized clinical trials. *Lifetime Data Anal.* **21**, 119–37.

Zhang, M., Tsiatis, A. A. & Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–15.

[*Received on* 28 *January* 2022. *Editorial decision on* 6 *July* 2023]