Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Log-rank and stratified log-rank tests

Ting Ye, Jun Shao & Yanyao Yi

To cite this article: Ting Ye, Jun Shao & Yanyao Yi (2023) Log-rank and stratified log-rank tests, Statistical Theory and Related Fields, 7:4, 309-317, DOI: 10.1080/24754269.2023.2263720

To link to this article: https://doi.org/10.1080/24754269.2023.2263720

STATISTICAL THEORY AND RELATED FIELDS
2023, VOL. 7, NO. 4, 309–317
https://doi.org/10.1080/24754269.2023.2263720

Log-rank and stratified log-rank tests

Ting Ye[a], Jun Shao[b,c] and Yanyao Yi[d]

[a]Department of Biostatistics, University of Washington, Seattle, WA, USA; [b]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [c]Department of Statistics, University of Wisconsin, Madison, WI, USA; [d]Global Statistical Sciences, Eli Lilly and Company, Indianapolis, IN, USA

ABSTRACT

In randomized clinical trials with right-censored time-to-event outcomes, the popular log-rank test without adjusting for baseline covariates is asymptotically valid for treatment effect under simple randomization of treatments but is too conservative under covariate-adaptive randomization. The stratified log-rank test, which adjusts baseline covariates in the test procedure by stratification, is asymptotically valid regardless of what treatment randomization is applied. In the literature, however, under simple randomization there is no affirmative conclusion about whether the stratified log-rank test is asymptotically more powerful than the unstratified log-rank test. In this article we show when the stratified and unstratified log-rank tests aim for the same null hypothesis and that, under simple randomization, the stratified log-rank test is asymptotically more powerful than the unstratified log-rank test in the region of alternative hypothesis that is specified by a Cox proportional hazards model. We also provide some discussion about why we do not have an affirmative conclusion in general.

ARTICLE HISTORY
Received 3 February 2023
Revised 24 June 2023
Accepted 20 September 2023

KEYWORDS
Baseline covariates; covariate-adaptive randomization; null hypothesis of no treatment effect; Pitman's relative efficiency; time-to-event; validity of tests

1. Introduction

The log-rank test (Mantel, 1966) and stratified log-rank test (Peto et al., 1976) are the two longstanding and most popular nonparametric tests for treatment effect in randomized clinical trials with two treatment arms and right-censored time-to-event outcomes. What motivates the stratified version of log-rank test is that baseline prognostic factors (covariates), measured prior to treatment assignments and thus not affected by treatments, are adjusted through stratification for efficiency gain.

Adjusting baseline covariates has been widely advocated to improve efficiency for tests and other analyzes, in the following two aspects. (i) In the design stage, covariate-adaptive randomization can be used to enforce the balance of treatment assignments across baseline prognostic factors, which results in more efficient tests (EMA, 2015). More details about covariate-adaptive randomization are given in Section 2. (ii) In the analysis stage, 'incorporating prognostic baseline factors in the primary statistical analysis of clinical trial data can result in a more efficient use of data to demonstrate and quantify the effects of treatment' (FDA, 2021), 'under approximately the same minimal statistical assumptions that would be needed for unadjusted' (EMA, 2015; FDA, 2021; ICH E9, 1998).

If the log-rank test is considered as 'unadjusted test', then the stratified log-rank test qualifies as an adjusted test under the same minimal assumption because it is still a nonparametric test without using any model. Tests using the Cox proportional hazards model as a working model are also qualified (DiRienzo & Lagakos, 2002; Kong & Slud, 1997; Lin & Wei, 1989), but the resulting tests can be less efficient than the unadjusted log-rank test when the working model is wrong (Kong & Slud, 1997). In this paper we focus on the stratified and unstratified log-rank tests.

Although stratified log-rank test uses information from baseline prognostic factors and thus is expected to be more efficient, an affirmative conclusion about whether it is asymptotically more efficient than the unstratified log-rank test is not available, under simple randomization in which patients are assigned to treatments completely at random. Another issue is that the stratified log-rank actually tests a null hypothesis stronger than that of the log-rank test and, hence, a prerequisite in their comparison is to investigate when the two null hypotheses are the same.

The purpose of this paper is to establish some affirmative conclusions about the stratified and unstratified log-rank tests, in terms of null hypothesis, asymptotic validity of tests and Pitman's asymptotic relative efficiency. The research is important as these two longstanding tests are used a lot in applications without a guidance on which one should be used.

CONTACT Jun Shao jshao@wisc.edu School of Statistics, East China Normal University, Shanghai 200241, People's Republic of China; Department of Statistics, University of Wisconsin, Madison, WI 53706, USA

Section 2 describes data, design, and log-rank test statistics. Section 3 introduces hypotheses, assumptions and the concept of validity for log-rank tests. Some theoretical results for stratified and unstratified log-rank tests are given in Section 4, where we show that, under simple randomization, the stratified log-rank test is asymptotically more powerful in the region of alternative hypothesis that is specified by a Cox proportional hazards model. Section 5 contains conclusions and Appendix provides technical proofs.

2. Data, design and test statistics

For a patient from the population under investigation, let T_j and C_j be the *potential* life time and right-censoring time, respectively, under treatment $j = 0$ or 1, and W be the vector of all baseline covariates and other time-varying covariates, observed or unobserved. 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 assigned and received.

Let I_i be a binary treatment indicator for patient i and $0 < \pi < 1$ be the pre-specified treatment assignment proportion for treatment 1. Consider the design, i.e. the generation of I_i's for n sequentially arrived patients. Simple randomization assigns patients to treatments completely at random with $P(I_i = 1) = \pi$ for all i, which may yield treatment proportions that substantially deviate from the target π across levels of some baseline prognostic factors. Because of this, covariate-adaptive randomization using Z, a sub-vector of W containing observed baseline prognostic factors with finitely many joint levels, is widely applied. When patient i with baseline $Z_i = z$ is arrived, a treatment is assigned using a mechanism dependent on all previously assigned treatments for patients with $Z_i = z$. For example, the most popular covariate-adaptive randomization scheme, the stratified permuted block design (Zelen, 1974), randomly assigns sequentially arrived patients with $Z_i = z$ in blocks of size B, each having $B\pi$ patients in treatment 1, where B is appropriately chosen so that $B\pi$ is an integer and the last block is allowed to be incomplete. Another popular covariate-adaptive randomization is Pocock-Simon's minimization (Pocock & Simon, 1975; Taves, 1974). Other schemes can be found in two reviews, Schulz and Grimes (2002) and Shao (2021). To see how popular covariate-adaptive randomization is, it was used in more than 500 clinical trials between 1989 and 2008 (Taves, 2010) and 237 trials among nearly 300 trials published in two years, 2009 and 2014 (Ciolino et al., 2019). All commonly used covariate-adaptive randomization schemes satisfy the following mild condition (Antognini & Zagoraiou, 2015).

(D1) Given $\{Z_1, \ldots, Z_n\}$, $\{I_1, \ldots, I_n\}$ and $\{T_{11}, C_{11}, T_{10}, C_{10}, W_1, \ldots, T_{n1}, C_{n1}, T_{n0}, C_{n0}, W_n\}$ are conditionally independent; $E(I_i \mid Z_1, \ldots, Z_n) = \pi$ 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 $I_i = 1$.

Most commonly used covariate-adaptive randomization schemes except Pocock-Simon's minimization also satisfy the next condition.

(D2) Conditional on $Z_1, \ldots, Z_n$, the vector whose zth component is $\sqrt{n}(n_{z1}/n - \pi)$ with z ranging over all levels of Z converges in distribution to $N(0, \Omega)$, where Ω is the diagonal matrix whose zth diagonal entry is $v/P(Z = z)$ and $v \le \pi(1 - \pi)$ is a known constant depending on the randomization scheme.

Although simple randomization is not counted as covariate-adaptive randomization, it satisfies (D1) and (D2) with $v = \pi(1 - \pi)$.

After I_i is assigned, the observed outcome from patient i is $\min(T_i, C_i)$ with $T_i = I_i T_{i1} + (1 - I_i)T_{i0}$ and $C_i = I_i C_{i1} + (1 - I_i)C_{i0}$, together with an indicator of $T_i \le C_i$.

The log-rank test statistic is

$$\begin{aligned} \top_{\mathrm{L}} &= \sqrt{n}\,\widehat{U}_{\mathrm{L}}/\widehat{\sigma}_{\mathrm{L}}, \\ \widehat{U}_{\mathrm{L}} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{I_i - \frac{\overline{Y}_1(t)}{\overline{Y}(t)}\right\}\mathrm{d}N_i(t), \\ \widehat{\sigma}^2_{\mathrm{L}} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\frac{\overline{Y}_1(t)\overline{Y}_0(t)}{\overline{Y}(t)^2}\,\mathrm{d}N_i(t), \end{aligned} \tag{1}$$

where $\overline{Y}(t) = \sum_{i=1}^n Y_i(t)/n$, $Y_i(t) = I_i Y_{i1}(t) + (1 - I_i)Y_{i0}(t)$, $Y_{ij}(t) =$ the indicator of the event $\min(T_{ij}, C_{ij}) \ge t$, $\overline{Y}_1(t) = \sum_{i=1}^n I_i Y_i(t)/n$, $\overline{Y}_0(t) = \sum_{i=1}^n (1 - I_i)Y_i(t)/n$, $N_i(t) = I_i N_{i1}(t) + (1 - I_i)N_{i0}(t)$, $N_{ij}(t)$ is the indicator of

the event $T_{ij} \le \min(t, C_{ij})$, and the upper limit τ in the integral is a point satisfying $P(\min(T_{ij}, C_{ij}) \ge \tau) > 0$ for $j = 0, 1$.

The stratified log-rank test statistic is a weighted average of the stratum-specific log-rank test statistics with strata constructed using Z,

$$\begin{aligned} \top_{\mathrm{SL}} &= \sqrt{n}\, \widehat{U}_{\mathrm{SL}} / \widehat{\sigma}_{\mathrm{SL}}, \\ \widehat{U}_{\mathrm{SL}} &= \frac{1}{n} \sum_{z} \sum_{i:Z_i=z} \int_0^{\tau} \left\{ I_i - \frac{\overline{Y}_{z1}(t)}{\overline{Y}_z(t)} \right\} \mathrm{d}N_i(t), \\ \widehat{\sigma}^2_{\mathrm{SL}} &= \frac{1}{n} \sum_{z} \sum_{i:Z_i=z} \int_0^{\tau} \frac{\overline{Y}_{z1}(t)\overline{Y}_{z0}(t)}{\overline{Y}_z(t)^2} \mathrm{d}N_i(t), \end{aligned} \tag{2}$$

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

It is clear that in terms of test statistics, the stratified $\top_{\mathrm{SL}}$ in (2) utilizes Z values whereas the unstratified $\top_{\mathrm{L}}$ in (1) is unadjusted. Under covariate-adaptive randomization, $\top_{\mathrm{L}}$ is not completely unadjusted since it uses Z-information through assignments I_i's, although it does not adjust for covariate-adaptive randomization in a correct way. On the other hand, the stratified $\top_{\mathrm{SL}}$ uses Z-information in both design and analysis stages.

We consider stratification with all levels of Z. In applications, it is allowed to use more covariates to form strata. The conclusions in what follows remain the same. However, it is not a good idea to use fewer levels of Z for stratification, because it may result in a test that is not asymptotically valid.

3. Null hypothesis, assumption and validity

Throughout, $\alpha \in (0, 1)$ denotes a given significance level and $z_{\alpha/2}$ is the $(1 - \alpha/2)$th quantile of the standard normal distribution. When $|\top_{\mathrm{L}}| > z_{\alpha/2}$, the log-rank test rejects the following null hypothesis H_0 of no treatment effect,

$$H_0 : \lambda_1(t) = \lambda_0(t) \quad \text{for all } t, \tag{3}$$

where $\lambda_j(t)$ is the unconditional hazard function of T_j, $j = 0, 1$. H_0 in (3) is a commonly adopted null hypothesis of no treatment effect unconditional on covariates.

The log-rank test is nonparametric. Its validity requires non-informative censoring (DiRienzo & Lagakos, 2002; Kong & Slud, 1997), i.e.,

(C) C_j is independent of T_j given j.

Under simple randomization, it is well-known (Kalbfleisch & Prentice, 2011) that the log-rank test is asymptotically valid in the sense that

$$\lim_{n\to\infty} P\left(|\top_{\mathrm{L}}| > z_{\alpha/2}\right) \le \alpha \tag{4}$$

with equality holding for at least one population P under H_0.

Unlike simple randomization, covariate-adaptive randomization generates a dependent sequence of treatment assignments, which may render conventional methods developed under simple randomization, such as the log-rank test, not valid under covariate-adaptive randomization (EMA, 2015; FDA, 2021). It is shown in Ye and Shao (2020) that, under covariate-adaptive randomization with ν in (D2) strictly smaller than $\pi(1 - \pi)$, the log-rank test is asymptotically conservative in the sense that,

$$\lim_{n\to\infty} P\left(|\top_{\mathrm{L}}| > z_{\alpha/2}\right) \le \alpha_0 < \alpha \tag{5}$$

for all P under H_0.

The stratified log-rank $\top_{\mathrm{SL}}$ in (2) actually tests the null hypothesis

$$\widetilde{H}_0 : \lambda_1(t \mid z) = \lambda_0(t \mid z) \quad \text{for all } t \text{ and } z, \tag{6}$$

where $\lambda_j(t \mid z)$ is the hazard function of T_j conditional on $Z = z$, $j = 0, 1$. Note that $\widetilde{H}_0$ in (6) holds if and only if the hazard functions are the same in every stratum z and, thus, is stronger than H_0 in (3).

The validity of stratified log-rank test requires the following assumption on censoring:

(CZ) C_j is independent of T_j given j and Z.

Conditions (C) and (CZ) are not comparable, although both are implied by that C_j is independent of (T_j, Z) given j, a reasonable condition for non-informative censoring.

Under simple randomization and covariate-adaptive randomization satisfying (D1) in Section 2, (4) holds with $\top_{\mathrm{L}}$ replaced by $\top_{\mathrm{SL}}$ and H_0 replaced by $\widetilde{H}_0$ (Ye & Shao, 2020), provided that all levels of Z are used in stratification.

Since $\widetilde{H}_0$ is stronger than H_0, the stratified and unstratified log-rank tests are not comparable. Thus, a prerequisite for the comparison of efficiency of two log-rank tests is $\widetilde{H}_0 = H_0$. Is there a scenario under which $\widetilde{H}_0 = H_0$? Consider the following transformation model assumption.

(TR) There is an increasing function h such that $h(P(T_0 \geq t \mid V)) = \theta + h(P(T_1 \geq t \mid V))$ for all (t, V) and a constant θ, where V is a vector of covariates, $Z \subset V \subset W$, and both h and θ can be unknown.

Assumption (TR) is discussed in Cheng et al. (1995), which includes many commonly used semiparametric models as special cases, for example, the Cox proportional hazards model (see formula (7) in Section 4). It is a mild assumption since h is unknown and we only need to know it exists.

The proof of following result is in the Appendix.

Theorem 3.1: *Under* (TR), $\widetilde{H}_0$ *in* (6) *is the same as* H_0 *in* (3).

4. Comparison of two log-rank tests

When $\widetilde{H}_0 = H_0$, is the stratified log-rank test $\top_{\mathrm{SL}}$ more efficient than the unstratified log-rank test $\top_{\mathrm{L}}$ under simple randomization when both tests are asymptotic valid? Intuitively this sounds correct since $\top_{\mathrm{L}}$ does not adjust for covariates.

Unfortunately, there is no result on this in the literature. In this section we try to fill this gap to some extent and explain why the two log-rank tests are not comparable in terms of efficiency. This is important because both stratified and unstratified log-rank tests are used a lot in applications.

To this goal, we first state the following asymptotic result (whose proof is given in Appendix) for the asymptotic distributions of stratified and unstratified log-rank tests under local alternatives. Define

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

where $\mu(t) = E(I_i \mid Y_i(t) = 1)$, $\mu_z(t) = E(I_i \mid Y_i(t) = 1, Z_i = z)$, $p(t)\,\mathrm{d}t = E\{\mathrm{d}N_i(t)\}/E\{Y_i(t)\}$, and $p_z(t)\,\mathrm{d}t = E\{\mathrm{d}N_i(t) \mid Z_i = z\}/E\{Y_i(t) \mid Z_i = z\}$. Also, we use O_j to denote O_{ij} for any i and O_{zj} to denote O_{zij} for any i and z. Note that, under the null hypothesis H_0, $E(O_j) = 0$ for $j = 0$, 1, and under the null hypothesis $\widetilde{H}_0$, $E(O_{zj} \mid Z = z) = 0$ for all z and $j = 0, 1$.

Theorem 4.1: (a) *Assume* (CZ) *and* (D1). *Under the local alternative hypothesis that* $E(O_{zj} \mid Z = z) = c_{zj}n^{-1/2}$ *with* c_{zj}*'s not depending on* n *and that* $\lambda_1(t \mid z)/\lambda_0(t \mid z)$ *is bounded and* $\to 1$ *for every* t *and* z, $\top_{\mathrm{SL}} \xrightarrow{d} N(\delta_{\mathrm{SL}}/\sigma_{\mathrm{SL}},\ 1)$, *where* $\xrightarrow{d}$ *denotes convergence in distribution as* $n \to \infty$, $\delta_{\mathrm{SL}} = \sum_z P(Z = z)\{\pi c_{z1} - (1-\pi)c_{z0}\}$, $\sigma^2_{\mathrm{SL}} = \sum_z P(Z = z)\{\pi \mathrm{var}_{\widetilde{H}_0}(O_{z1} \mid Z = z) + (1-\pi)\mathrm{var}_{\widetilde{H}_0}(O_{z0} \mid Z = z)\}$, *and* $\mathrm{var}_{\widetilde{H}_0}$ *denotes variance under* $\widetilde{H}_0$.

(b) *Assume* (C), (D1), *and* (D2). *Under the local alternative hypothesis that* $E(O_{zi}) = c_j n^{-1/2}$ *with* c_j*'s not depending on* n *and that* $\lambda_1(t)/\lambda_0(t)$ *is bounded and* $\to 1$ *for every* t, $\top_{\mathrm{L}} \xrightarrow{d} N(\delta_{3\mathrm{L}}/\sigma_{\mathrm{L}},\ \sigma^2_{\mathrm{L}}(\nu)/\sigma^2_{\mathrm{L}})$, *where* $\delta_{\mathrm{L}} = \pi c_1 - (1-\pi)c_0$, $\sigma^2_{\mathrm{L}} = \pi \mathrm{var}_{H_0}(O_1) + (1-\pi)\mathrm{var}_{H_0}(O_0)$, $\sigma^2_{\mathrm{L}}(\nu) = \sigma^2_{\mathrm{L}} - \{\pi(1-\pi) - \nu\}\mathrm{var}_{H_0}\{E_{H_0}(O_1|Z) + E_{H_0}(O_0|Z)\}$ *for* ν *given in (D2), and* E_{H_0} *and* var_{H_0} *denote expectation and variance under* H_0, *respectively.*

Because the local alternative hypotheses specified in (a) and (b) of Theorem 4.1 do not follow any model, δ^2_{SL} and δ^2_{L} can be arbitrarily very different and, thus, $\top_{\mathrm{SL}}$ and $\top_{\mathrm{L}}$ may be not comparable in terms of asymptotic efficiency. In other words, the space of alternative hypothesis is too large to compare efficiency of $\top_{\mathrm{SL}}$ and $\top_{\mathrm{L}}$, as there is no model at all. A semiparametric model on alternative hypothesis narrowing down the space of alternative hypothesis may result in affirmative results of comparing efficiency. We derive a result under the Cox proportional hazards model to highlight this.

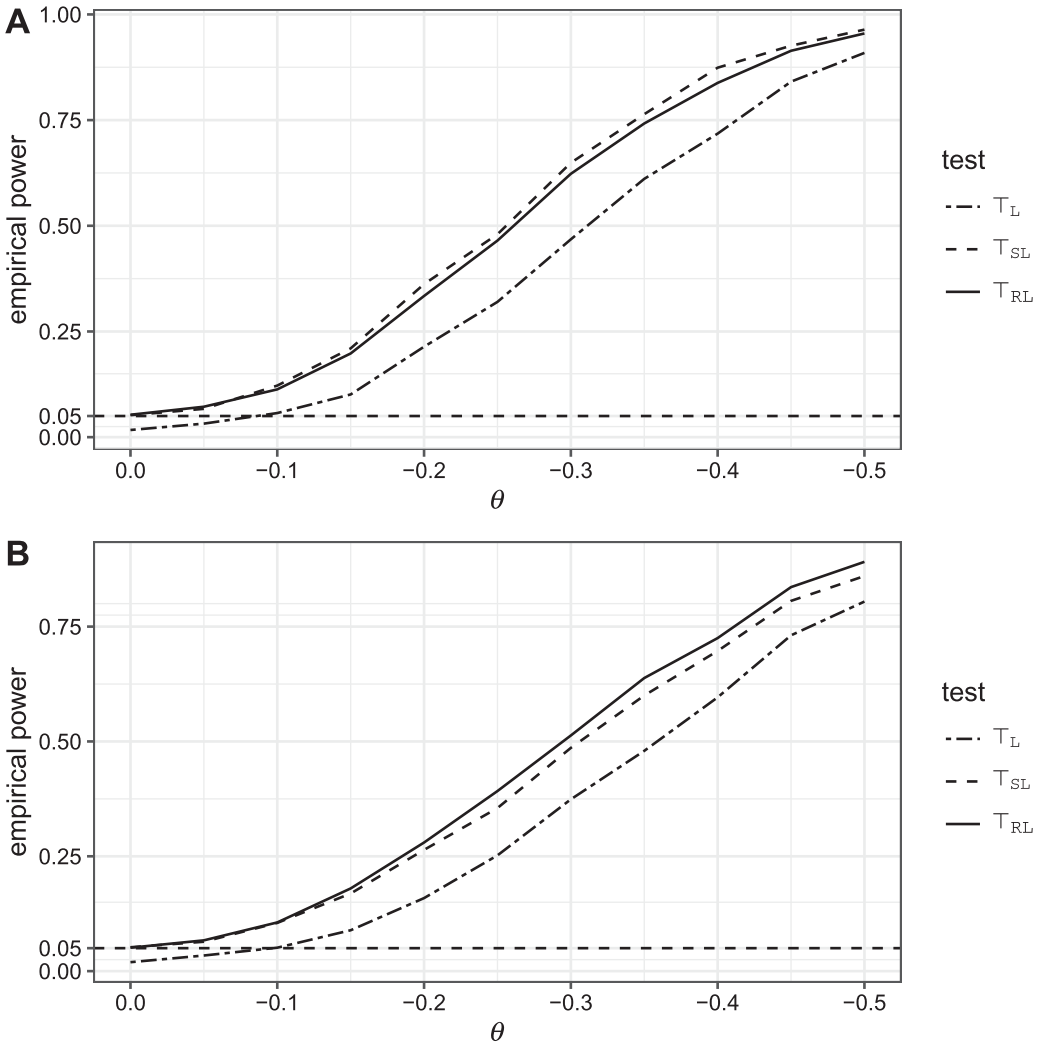

Figure 1. Power curves based on $n = 500$ and 2000 simulations.

Suppose that the true hazard follows a Cox proportional hazards model,

$$\lambda_j(t \mid V) = \lambda(t)\exp(\theta j + \eta^\top V), \quad j = 0, 1, \tag{7}$$

where $Z \subset V \subset W, \lambda_j(t \mid V)$ is the hazard conditional on covariate V, θ is an unknown parameter, η is an unknown parameter vector, and $\lambda(t)$ is an unspecified function. Under model (7), (TR) holds with $h(s) = -\log(-\log(s))$ and $\widetilde{H}_0 = H_0 : \theta = 0$.

Corollary 4.1: *Assume that model* (7) *holds, C_j is independent of (T_j, Z) given j, and $P(C_1 \geq t \mid V) = P(C_0 \geq t \mid V)$ for all t. Then, under simple randomization, the stratified log-rank test $\top_{SL}$ is always more efficient than the unstratified log-rank test $\top_L$ in terms of Pitman's asymptotic relative efficiency.*

The proof is given in the Appendix. A key to the proof is that the local alternative hypotheses in (a) and (b) of Theorem 4.1 can be unified into $\theta = c/\sqrt{n}$ with the help of model (7).

As both log-rank tests are nonparametric and do not need model (7), what does Corollary 4.1 tell us? It says that, under simple randomization, the stratified log-rank test $\top_{SL}$ is more efficient in the region of alternative hypothesis specified by model (7), although we cannot claim that $\top_{SL}$ is more efficient in the entire alternative hypothesis space.

We now turn to covariate-adaptive randomization, under which the unstratified log-rank test $\top_L$ is not valid but conservative, as we discussed in Section 3. On the other hand, by Theorem 4.1(a), the stratified log-rank test $\top_{SL}$ is valid for testing $\widetilde{H}_0$ regardless of which covariate-adaptive randomization is applied. Therefore, stratified log-rank test is a clear winner when covariate-adaptive randomization is applied.

Another way to adjust for covariates used in randomization is the modified (unstratified) log-rank test $\top_{RL} = \widehat{\sigma}_L \top_L / \widehat{\sigma}_L(\nu)$ proposed by Ye and Shao (2020), where $\widehat{\sigma}_L(\nu)$ is a consistent estimator of $\sigma_L(\nu)$ (see §3.2 of Ye and Shao 2020). $\top_{RL}$ removes the conservativeness of $\top_L$ and is valid for testing H_0 in (3) under covariate-adaptive randomization.

Even if model (7) holds, $\top_{\mathrm{SL}}$ and $\top_{\mathrm{RL}}$ are not comparable in terms of asymptotic efficiency. We provide two simulation examples here to demonstrate that $\top_{\mathrm{SL}}$ is more efficient in one scenario but less efficient in another scenario, compared with $\top_{\mathrm{RL}}$. The simulation setting is model (7) with $\lambda(t) = 12^{-1}\log 2$ for all t and $\eta^\top V = -1.5Z_1 + 0.5Z_2^2$, where Z_1 is binary with $P(Z_1 = 1) = 0.5$, $Z_2 \sim N(0,1)$, and Z_1 and Z_2 are independent. Z_1 and discretized Z_2 with 4 equal probability categories are used for stratified permuted block randomization with block size 4. In scenario 1, censoring is independent of treatment and (Z_1, Z_2) and distributed as uniform on (10,40). In scenario 2, censoring is independent of treatment and Z_2, but conditioned on Z_1, and censoring is distributed as 10 + the exponential distribution with mean $2Z_1$. The power curves over θ with $\alpha = 0.05$ and $n = 500$ based on 2000 simulations are given in Figure 1. Note that $\top_{\mathrm{SL}}$ is more powerful than $\top_{\mathrm{RL}}$ under scenario 1 but less powerful under scenario 2. Both $\top_{\mathrm{SL}}$ and $\top_{\mathrm{RL}}$ are more powerful than the conservative $\top_{\mathrm{L}}$ in any case.

The reason why the stratified $\top_{\mathrm{SL}}$ and the modified unstratified $\top_{\mathrm{RL}}$ are not comparable in asymptotic efficiency is that the two tests adopt different approaches in utilizing baseline covariates: the former adjusts baseline covariates by stratification, whereas the latter utilizes baseline covariates by modifying the unstratified $\top_{\mathrm{L}}$ whose performance is affected by covariate-adaptive randomization.

5. Conclusion and discussion

(1) Under some semiparametric models for survival time such as the transformation model (TR) described in Section 3, the null hypotheses of stratified and unstratified log-rank tests are the same.
(2) Under simple randomization of treatment assignments, the stratified log-rank test is asymptotically more efficient than the unstratified log-rank test in terms of Pitman's relative efficiency in the region of alternative hypothesis specified by the Cox proportional hazards model given by (7). It is of interest to derive more affirmative results using assumptions/models other than the Cox model to narrow down the space of alternative hypothesis.
(3) Under covariate-adaptive randomization of treatment assignments, the unstratified log-rank test is not asymptotically valid but conservative, whereas the stratified log-rank test is asymptotically valid as long as the covariates used in randomization are all included in stratification. Thus, the stratified log-rank test is a clear winner. A modified unstratified log-rank test removes conservativeness and is valid, but its relative efficiency compared with the stratified log-rank test has no definite conclusion, because the two tests apply different approaches in utilizing covariates.
(4) Because the region specified by the Cox model is quite large and the stratified log-rank test is a clear winner under covariate-adaptive randomization, we recommend the stratified log-rank test over the unstratified log-rank test.

Acknowledgements

We would like to thank two anonymous referees and an associate editor for helpful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Antognini, A. B., & Zagoraiou, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli Journal*, *21*(2), 881–908.

Cheng, S., Wei, L., & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, *82*(4), 835–845. https://doi.org/10.1093/biomet/82.4.835

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 Medical Research Methodology*, *19*(1), 136. https://doi.org/10.1186/s12874-019-0787-8

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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(4), 745–757. https://doi.org/10.1111/1467-9868.00310

EMA (2015). *Guideline on adjustment for baseline covariates in clinical trials*. Committee for Medicinal Products for Human Use, European Medicines Agency (EMA).

FDA (2021). *Adjusting for covariates in randomized clinical trials for drugs and biological products*. Draft 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. May 2021.

ICH E9 (1998). *Statistical principles for clinical trials E9*. International Council for Harmonisation (ICH).

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data*. Wiley.

Kong, F. H., & Slud, E. (1997). Robust covariate-adjusted logrank tests. *Biometrika*, *84*(4), 847–862. https://doi.org/10.1093/biomet/84.4.847
Lin, D. Y., & Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, *84*(408), 1074–1078. https://doi.org/10.1080/01621459.1989.10478874
Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, *50*(3), 163–170.
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. *British Journal of Cancer*, *34*(6), 585–612. https://doi.org/10.1038/bjc.1976.220
Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, *31*(1), 103–115. https://doi.org/10.2307/2529712
Schulz, K. F., & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: Chance, not choice. *The Lancet*, *359*(9305), 515–519. https://doi.org/10.1016/S0140-6736(02)07683-3
Shao, J. (2021). Inference for covariate-adaptive randomization: Aspects of methodology and theory (with discussions). *Statistical Theory and Related Fields*, *5*(3), 172–186. https://doi.org/10.1080/24754269.2021.1871873
Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, *15*(5), 443–453. https://doi.org/10.1002/cpt.1974.15.issue-5
Taves, D. R. (2010). The use of minimization in clinical trials. *Contemporary Clinical Trials*, *31*(2), 180–184. https://doi.org/10.1016/j.cct.2009.12.005
Ye, T., & Shao, J. (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(5), 1301–1323. https://doi.org/10.1111/rssb.12392
Ye, T., Shao, J., Yi, Y., & Zhao, Q. (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*.
Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, *27*(7-8), 365–375. https://doi.org/10.1016/0021-9681(74)90015-0

Appendix

Proof of Theorem 3.1

It is clear that $\widetilde{H}_0$ in (6) implies H_0 in (3). Thus, it suffices to show that, under (TR), $\lambda_1(t) = \lambda_0(t)$ for all t implies $\lambda_1(t \mid Z) = \lambda_0(t \mid Z)$ for all (t, Z). Define $S_j(t \mid V) = P(T_j \geq t \mid V)$ and $S_j(t) = P(T_j \geq t)$. If $\lambda_1(t) = \lambda_0(t)$ for all t, then $S_0(t) = S_1(t)$ for all t. Condition (TR) implies that

$$S_0(t \mid V) = h^{-1}(\theta + h(S_1(t \mid V))) \quad \text{for all } (t, V).$$

Then

$$S_1(t) = S_0(t) = E\{S_0(t \mid V)\} = E\left\{h^{-1}(\theta + h(S_1(t \mid V)))\right\} \quad \text{for all } t,$$

i.e.,

$$E\left\{h^{-1}(\theta + h(S_1(t \mid V))) - S_1(t \mid V)\right\} = 0 \quad \text{for all } t.$$

Since $h^{-1}(\theta + h(S_1(t \mid V))) \geq S_1(t \mid V)$ or $\leq S_1(t \mid V)$ depending on whether $\theta \geq 0$ or ≤ 0,

$$h^{-1}(\theta + h(S_1(t \mid V))) = S_1(t \mid V) \quad \text{for all } (t, V).$$

This implies that $\theta = 0$ and, thus, $S_0(t \mid V) = S_1(t \mid V)$ for all (t, V), which together with $Z \subset V$ imply that $S_0(t \mid Z) = S_1(t \mid Z)$ and hence $\lambda_0(t \mid Z) = \lambda_1(t \mid Z)$ for all (t, Z).

Proof of Theorem 4.1

We prove (a) only, since the proof of (b) is similar. We first show that, under the null hypothesis $\widetilde{H}_0$ or alternative hypothesis,

$$\sqrt{n}\left(\widehat{U}_{\mathrm{SL}} - \sum_z \frac{n_{z1}\theta_{z1} - n_{z0}\theta_{z0}}{n}\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}^2_{\mathrm{SL}}\right), \tag{A1}$$

where n_{zj} = the number of patients with treatment j in stratum z, $\theta_{zj} = E(O_{zj} \mid Z = z)$, $j = 0, 1$, and

$$\widetilde{\sigma}^2_{\mathrm{SL}} = \sum_z P(Z = z)\{\pi \mathrm{var}(O_{z1} \mid Z = z) + (1 - \pi)\mathrm{var}(O_{z0} \mid Z = z)\}.$$

Following the argument in the Appendix of Lin and Wei (1989), we obtain that, under either the null or alternative hypothesis, the left hand side of (A1) is equal to

$$\frac{1}{\sqrt{n}} \sum_z \sum_{i:Z_i = z} \{I_i(O_{zi1} - \theta_{z1}) - (1 - I_i)(O_{zi0} - \theta_{z0})\} + o_p(1), \tag{A2}$$

where $o_p(1)$ denotes a quantity converging to 0 in probability as $n \to \infty$. Define $\mathcal{I} = \{I_1, \ldots, I_n\}$ and $\mathcal{Z} = \{Z_1, \ldots, Z_n\}$. Similar to the proof of Theorem 2 in Ye et al. (2022), the Lindeberg's Central Limit Theorem justifies that, conditioned on $\mathcal{I}$ and $\mathcal{Z}$, the

random vector

$$\left(\frac{1}{\sqrt{n}} \sum_z \sum_{i:Z_i=z} I_i(O_{zi1} - \theta_{z1}),\ \frac{1}{\sqrt{n}} \sum_z \sum_{i:Z_i=z} (1-I_i)(O_{zi0} - \theta_{z0}) \right)^\top$$

converges in distribution to a 2-dimensional normal distribution with mean 0, conditional on $\mathcal{I}$ and $\mathcal{Z}$. Let M be the quantity in (A2) excluding $o_p(1)$, which is the sum of two components of the previous random vector. Consequently,

$$\{\text{var}(M \mid \mathcal{I}, \mathcal{Z})\}^{-1/2} M \mid \mathcal{I}, \mathcal{Z} \xrightarrow{d} N(0,1).$$

Under (D1),

$$\begin{aligned}
\text{var}(M \mid \mathcal{I}, \mathcal{Z}) &= \frac{1}{n} \sum_z \left\{ \sum_{i:Z_i=z, I_i=1} \text{var}(O_{iz1} \mid \mathcal{Z}) + \sum_{i:Z_i=z, I_i=0} \text{var}(O_{iz0} \mid \mathcal{Z}) \right\} \\
&= \frac{1}{n} \sum_z \left\{ \sum_{i:Z_i=z, I_i=1} \text{var}(O_{iz1} \mid Z_i = z) + \sum_{i:Z_i=z, I_i=0} \text{var}(O_{iz0} \mid Z_i = z) \right\} \\
&= \frac{1}{n} \sum_z \{ n_{z1} \text{var}(O_{z1} \mid Z = z) + n_{z0} \text{var}(O_{z0} \mid Z = z) \} \\
&= \sum_z \frac{n_z}{n} \left\{ \frac{n_{z1}}{n_z} \text{var}(O_{z1} \mid Z = z) + \frac{n_{z0}}{n_z} \text{var}(O_{z0} \mid Z = z) \right\} \\
&= \sum_z P(Z = z) \{ \pi \, \text{var}(O_{z1} \mid Z = z) + (1-\pi) \text{var}(O_{z0} \mid Z = z) \} + o_p(1) \\
&= \widetilde{\sigma}^2_{\text{SL}} + o_p(1).
\end{aligned}$$

Then $\{\text{var}(M \mid \mathcal{I}, \mathcal{Z})\}^{-1/2} M \xrightarrow{d} N(0,1)$ unconditionally. Thus, by Slutsky's theorem, (A1) holds.

Next, under the local alternative specified in part (a), $\widetilde{\sigma}^2_{\text{SL}} \to \sigma^2_{\text{SL}}$ and

$$\begin{aligned}
\sqrt{n} \left(\sum_z \frac{n_{z1}\theta_{z1} - n_{z0}\theta_{z0}}{n} \right) &= \sum_z P(Z = z) \{ \pi \, c_{z1} - (1-\pi) c_{z0} \} + o_p(1) \\
&= \delta_{\text{SL}} + o_p(1).
\end{aligned}$$

Hence, by (A1) and Slutsky's theorem,

$$\sqrt{n} \widehat{U}_{\text{SL}} \xrightarrow{d} N\left(\delta_{\text{SL}}, \sigma^2_{\text{SL}} \right).$$

It remains to show that $\widehat{\sigma}^2_{\text{SL}} - \sigma^2_{\text{SL}} = o_p(1)$, under the specified local alternative. By Lemma 3 of Ye and Shao (2020), within any stratum z, $\overline{Y}_{z1}(t)\overline{Y}_{z0}(t)/\overline{Y}_z(t)^2 = \mu_z(t)\{1 - \mu_z(t)\} + o_p(1)$. By the identity

$$E\{\text{d}N_i(t)\} = \pi E\{Y_{i1}(t)\lambda_1(t)\} \, \text{d}t + (1-\pi) E\{Y_{i0}(t)\lambda_0(t)\} \, \text{d}t$$

from Kalbfleisch and Prentice (2011) and the form of $\widehat{\sigma}^2_{\text{SL}}$, we obtain that, under the specified local alternative,

$$\begin{aligned}
\widehat{\sigma}^2_{\text{SL}} &= \sum_z \frac{n_z}{n} \int_0^\tau \mu_z(t)\{1 - \mu_z(t)\} E\{\text{d}N_i(t)\} + o_p(1) \\
&= \sum_z \frac{n_z}{n} \int_0^\tau \mu_z(t)\{1 - \mu_z(t)\} [\pi E\{Y_{i1}(t)\lambda_1(t)\} + (1-\pi) E\{Y_{i0}(t)\lambda_0(t)\}] \, \text{d}t + o_p(1) \\
&= \sum_z P(Z = z) \{ \pi \, \text{var}_{\widetilde{H}_0}(O_{z1} \mid Z = z) + (1-\pi) \text{var}_{\widetilde{H}_0}(O_{z0} \mid Z = z) \} + o_p(1) \\
&= \sigma^2_{\text{SL}} + o_p(1).
\end{aligned}$$

Proof of Corollary 4.1

A direct calculation shows that

$$\sigma^2_{\text{L}} = \sigma^2_{\text{SL}} = \int_0^\tau E_{H_0}\{Y_i(t) \exp(\eta^\top V_i)\} \nu(t) \, \text{d}\Lambda(t),$$

where σ^2_{L} and σ^2_{SL} are given in Theorem 4.1, $\Lambda(t) = \int_0^t \lambda(s) \, \text{d}s$, $\nu(t) = \text{var}(I_i \mid Y_i(t) = 1)$, and E_{H_0} denotes expectation under $H_0 : \theta = 0$.

Under the local alternative hypothesis $\theta = c/\sqrt{n}$ with a fixed constant $c \neq 0$, by Theorem 4.1, $\top_{\mathrm{L}} \xrightarrow{d} N(c\,\theta_{\mathrm{L}}/\sigma_{\mathrm{L}},\, 1)$, where

$$\theta_{\mathrm{L}} = \sigma_{\mathrm{L}}^2 - \int_0^\tau \left(E_{H_0}\left\{ Y_i(t) \exp(2\eta^\top V_i) \right\} - \frac{\left[E_{H_0}\{ Y_i(t) \exp(\eta^\top V_i) \} \right]^2}{E_{H_0}\{ Y_i(t) \}} \right) \nu(t) \Lambda(t)\, \mathrm{d}\Lambda(t),$$

and $\top_{\mathrm{SL}} \xrightarrow{d} N(c\,\theta_{\mathrm{SL}}/\sigma_{\mathrm{L}},\, 1)$, where

$$\theta_{\mathrm{SL}} = \sigma_{\mathrm{L}}^2 - \int_0^\tau \left(E_{H_0}\left\{ Y_i(t) \exp(2\eta^\top V_i) \right\} - E_{H_0} \frac{\left[E\{ Y_i(t) \exp(\eta^\top V_i) \mid Z_i \} \right]^2}{E_{H_0}\{ Y_i(t) \mid Z_i \}} \right) \nu(t) \Lambda(t)\, \mathrm{d}\Lambda(t).$$

Pitman's asymptotic relative efficiency of $\top_{\mathrm{SL}}$ with respect to $\top_{\mathrm{L}}$ is $\theta_{\mathrm{SL}}^2/\theta_{\mathrm{L}}^2$.

Applying Jensen's inequality $\varphi\{E(M)\} \leq E\{\varphi(M)\}$ with convex function $\varphi(t_1, t_2) = t_1^2/t_2$ and $M = (E_{H_0}\{ Y_i(t) \exp(\eta^\top V_i) \} \mid Z_i\}, E_{H_0}\{ Y_i(t) \mid Z_i \})^\top$, we obtain that $\theta_{\mathrm{L}} \leq \theta_{\mathrm{SL}}$. To reach the conclusion $\theta_{\mathrm{SL}}^2/\theta_{\mathrm{L}}^2 \geq 1$, it remains to show that $\theta_{\mathrm{L}} \geq 0$.

The condition $P(C_1 \geq t \mid V) = P(C_0 \geq t \mid V)$ for all t implies that $\nu(t) = \pi(1-\pi)$ and, hence,

$$\begin{aligned}
\theta_{\mathrm{L}} = {} & \pi(1-\pi) \int_0^\tau E_{H_0}\left\{ Y_i(t) \exp(\eta^\top V_i) \right\} \mathrm{d}\Lambda(t) \\
& - \pi(1-\pi) \int_0^\tau E_{H_0}\left\{ Y_i(t) \exp(2\eta^\top V_i) \right\} \Lambda(t)\, \mathrm{d}\Lambda(t) \\
& + \pi(1-\pi) \int_0^\tau \frac{\left[E_{H_0}\{ Y_i(t) \exp(\eta^\top V_i) \} \right]^2}{E_{H_0}\{ Y_i(t) \}} \Lambda(t)\, \mathrm{d}\Lambda(t).
\end{aligned}$$

Thus, it suffices to show

$$\int_0^\tau \left[E_{H_0}\left\{ Y_i(t) \exp(\eta^\top V_i) \right\} - E_{H_0}\left\{ Y_i(t) \exp(2\eta^\top V_i) \right\} \Lambda(t) \right] \mathrm{d}\Lambda(t) \geq 0. \tag{A3}$$

Note that

$$\begin{aligned}
E\{N_{i1}(\tau)\} &= \int_0^\tau E\{\mathrm{d}N_{i1}(t)\} \\
&= \int_0^\tau E\left\{ Y_{i1}(t) \exp(\theta + \eta^\top V_i) \right\} \mathrm{d}\Lambda(t) \\
&= \int_0^\tau E_V \left[\exp\{-\Lambda(t) \exp(\theta + \eta^\top V_i)\} P(C_1 \geq t \mid V_i) \exp(\theta + \eta^\top V_i) \right] \mathrm{d}t,
\end{aligned}$$

where E_V is the expectation with respect to covariate V_i and is not depending on θ. Taking the derivative with respect to θ, we obtain that

$$\begin{aligned}
\frac{\partial E\{N_{i1}(\tau)\}}{\partial \theta} = \int_0^\tau E_V \Big[& \exp\{-\Lambda(t) \exp(\theta + \eta^\top V_i)\} P(C_1 \geq t \mid V_i) \exp(\theta + \eta^\top V_i) \\
& - \exp\{-\Lambda(t) \exp(\theta + \eta^\top V_i)\} P(C_1 \geq t \mid V_i) \exp(2\theta + 2\eta^\top V_i) \Lambda(t) \Big] \mathrm{d}\Lambda(t).
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{\partial E\{N_{i1}(\tau)\}}{\partial \theta} |_{\theta=0} = \int_0^\tau E_V \Big[& \exp\{-\Lambda(t) \exp(\eta^\top V_i)\} P(C_1 \geq t \mid V_i) \exp(\eta^\top V_i) \\
& - \exp\{-\Lambda(t) \exp(\eta^\top V_i)\} P(C_1 \geq t \mid V_i) \exp(2\eta^\top V_i) \Lambda(t) \Big] \mathrm{d}\Lambda(t) \\
= \int_0^\tau & \left[E_{H_0}\left\{ Y_{i1}(t) \exp(\eta^\top V_i) \right\} - E_{H_0}\left\{ Y_{i1}(t) \exp(2\eta^\top V_i) \Lambda(t) \right\} \right] \mathrm{d}\Lambda(t),
\end{aligned}$$

which is the same as the left-hand side of (A3). As $E\{N_{i1}(\tau)\}$ is the probability of having an observed failure before time τ, it is a non-decreasing function of θ. This implies that (A3) holds.