Statistica Sinica Preprint No: SS-2022-0011	
Title	Testing Hypotheses of Covariate-Adaptive Randomized Clinical Trials with Time-to-event Outcomes under the AFT Model
Manuscript ID	SS-2022-0011
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0011
Complete List of Authors	Hongjian Zhu, Lixin Zhang, Jing Ning and Lu Wang
Corresponding Authors	Lixin Zhang
E-mails	stazlx@zju.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Testing Hypotheses of Covariate-Adaptive Randomized Clinical Trials with Time-to-event Outcomes under the AFT model

Hongjian Zhu[1], Li-Xin Zhang[2*], Jing Ning[3] and Lu Wang[4]

[1] *AbbVie Inc.*

[2*] *Zhejiang Gongshang University and Zhejiang University*

[3] *University of Texas MD Anderson Cancer Center*

[4] *University of Texas Health Science Center at Houston*

Abstract: Covariate adaptive randomization (CAR) designs, including the stratified permuted block randomization design, are popular in clinical trials. However, clinical trialists usually ignore the unique feature of the CAR that the treatment assignment of the current subject depends not only on his or her covariate information, but also on the covariates and treatment assignments of all prior subjects. They often analyze the data as if complete randomization was used. As a result, the inferential conclusions of many clinical trials are open to question. This paper provides the theoretical foundation for trials using CAR designs and the accelerated failure time (AFT) model for time-to-event outcomes. We derive the asymptotic properties of the test statistics and explain the effect of the CAR design on the variability of the estimated treatment effect and the type

*Email:stazlx@zju.edu.cn

I error rate. We also obtain the consistency and asymptotic normality of the estimators. Based on the theoretical results, we propose new test statistics to control the type I error rate. Numerical studies demonstrate our theoretical findings and show that our methods successfully protect the type I error rate. Our theoretical and numerical results provide practical guidance for future clinical trials employing CAR designs and time-to-event outcomes.

Key words and phrases: Accelerated failure time model, Conservative tests, Covariate adaptive design, Type I error.

1. Introduction

Covariate adaptive randomization (CAR) is popular in clinical trials (Rosenberger and Lachin, 2015) and development economics research (Duflo et al., 2007; Bruhn and McKenzie, 2009). However, the validity of the inference following CAR has been questioned (Weir and Lees, 2003; Hagino et al., 2004). In this paper, we study the validity of inference in clinical trials with CAR and the accelerated failure time (AFT) model for time-to-event outcomes.

We first introduce the importance and advantages of CAR designs. It is well accepted that many covariates (biomarkers) are associated with certain diseases (Ashley et al., 2010), and this has led to precision medicine. When designing an efficient clinical study for precision medicine, clinical trial practitioners' first and most common concern is balancing the treat-

ment allocation for influential covariates. CAR design balances the patients' prognostic factors in each treatment arm by sequentially assigning the next patient to an arm based on the current covariate and the previous treatment assignments and covariates. CAR design can avoid the inaccuracy introduced into the estimation of treatment effects by a poor balance in the patients' characteristics. A balance in prognostic factors across treatments is also desirable for clinical trials that, for example, have a small sample size, involve interim analysis, or require subgroup analysis (Toorawa et al., 2009). An overview of CAR designs can be found in Rosenberger and Sverdlov (2008). Stratified permuted block randomization (Taves, 1974) is popular not only in clinical trials but also in economic research (Bugni et al., 2018), and from 1989 to 2008 Pocock and Simons marginal procedure (Pocock and Simon, 1975) was used in over 500 trials (Taves, 2010). Other CAR designs can be found in Nordle and Brantmark (1977), Wei (1978), Signorini et al. (1993), Heritier et al. (2005), Russell et al. (2011), Hu and Hu (2012), Lebowitsch et al. (2012), and Antognini and Zagoraiou (2011).

Next, we briefly discuss a major issue with CAR designs. The Student's t-test is common in clinical trials (Sverdlov, 2015), and only between 24% and 34% of randomized trials adjust covariates in their main analyses (Kahan et al., 2014). There are several reasons for not using the full model

to adjust covariates. It is difficult to incorporate some covariates into the working model; for example, in a multi-center trial the investigation site is usually omitted from the analysis. Furthermore, fewer covariates improve the simplicity and transparency of the test procedure. Adjusting too many covariates usually leads to a more complicated model that is less robust to model misspecification. However, through simulation, researchers have realized that CAR will lead to a conservative type I error rate if some of the randomization covariates are omitted from the analysis (see, e.g., Birkett (1985), Forsythe (1987), Aickin (2002), Weir and Lees (2003), Hagino et al. (2004)). Shao et al. (2010), Ma et al. (2015, 2020), and Shao and Yu (2013) have offered a theoretical explanation for such conservativeness in the context of linear regression models and generalized linear models. Further, Bugni et al. (2018) investigated robust inference under CAR.

We study the validity of the statistical inference and the control of the type I error rate for clinical trials with CAR and the AFT model for time-to-event outcomes. The AFT model is an essential alternative to the proportional hazards model. For example, in the CREST trial (Lal et al., 2012), the stratified permuted block design was used to balance allocation over two characteristics (center and symptomatic status), and the AFT model was fitted to study treatment differences in the restenosis rates. There-

fore, it is significant to investigate clinical trials with CAR and AFT due to both approaches' advantages and importance. Inference procedures for AFT and their asymptotic properties have been studied extensively (Buckley and James (1979), Koul et al. (1981), Tsiatis (1990), Miller and Halpern (1982), Ritov (1990), Lai and Ying (1991), Ying (1993), Lin and Ying (1995), Jin et al. (2003), Leon et al. (2009), Stute (1993, 1996)). However, the validity of inference under the CAR and AFT models has not been explicitly investigated.

The difficulties of our research include the complicated correlation structure of the within-stratum imbalances and the allocation probability function's discreteness. Moreover, the dependence among the survival times, the covariates, and the assignments complicates the study of the properties of the estimators of both the survival function and the parameters. For example, in linear or generalized linear models the covariates in the estimator of the parameter can be separated into balanced and unbalanced parts; but the estimator in the AFT model depends on an estimator of a non-parameter function, so the covariates cannot be separated directly. Additional challenges include the incomplete data due to right censoring and the variability caused by inverse provability weighting. We use advanced theoretical techniques such as martingale theory to overcome these

challenges, and we study a general family of CAR designs. Note that (i) we consider the q-balance CAR to satisfy various requirements of clinical trials; (ii) we study a family of CAR designs that include not only SPB, Pocock and Simon's design, but also that of Hu and Hu (2012). Under some conditions, we obtain the fundamental theory for inference for trials with CAR and AFT, including the asymptotic properties of the test statistics and the consistency and asymptotic normality of the parameter estimators. As a result, the type I error rate can be well controlled. In addition to the main treatment effects, we investigate the general form of hypothesis testing for the significance of the covariates. Further important lemmas and theoretical conclusions can be found in the Supplementary Material.

We show that CAR procedures will shrink the variability of the estimated treatment effect and give a conservative type I error rate if we do not include all the randomization covariates in the data analysis. We propose model-based approaches to adjust the estimated variance of the treatment effects, and the numerical results show that our methods successfully protect the type I error rate.

The remainder of our article is organized as follows. In Section 2, we introduce the framework and the major theorems. We give the numerical results in Section 3 and provide a discussion in Section 4. The proofs are

given in the Supplementary Material.

2. Statistical Inference in Clinical Trials with Survival Analysis and CAR Designs

2.1 Framework

Consider a randomized clinical trial in which CAR designs are used to sequentially assign n subjects to one of two treatments. Let I_i, $i = 1, 2, \ldots, n$, indicate the assignment of the ith patient, i.e., $I_i = 1$ for treatment 1 and $I_i = 0$ for treatment 2. Let T_i be the survival times and C_i be the censoring times. Denote the covariates of interest by $\boldsymbol{X}_i = (X_{i,1}, X_{i,2} \ldots X_{i,p_1})^T$ and $\boldsymbol{Z}_i = (Z_{i,1}, Z_{i,2} \ldots Z_{i,p_2})^T$, and assume that the CAR designs are applied with respect to both the $\boldsymbol{X}_i$s and $\boldsymbol{Z}_i$s, but only the $\boldsymbol{X}_i$s are used in the analysis. More details of the CAR design will be given in Section 2.2. The observed data are represented by $(Y_i, \delta_i, \boldsymbol{X}_i, \boldsymbol{Z}_i), i = 1, 2, ..., n$, where $Y_i = min(T_i, C_i), \delta_i = I\{T_i \leq C_i\}$, and $I\{\cdot\}$ is the indicator function. Following the assumptions of Cheng et al. (1995) and Shen et al. (2009), we assume that the survival function $G(\cdot)$ of C_i does not depend on $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ and $F_G(\cdot) = 1 - G(\cdot)$ is the cumulative distribution of C_i. Further, the censoring time C_i and the survival time T_i are assumed to be independent given the covariates $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$.

Recall that the AFT model gives a linear relationship between the logarithms of the survival time and the covariates of interest (Kalbfleisch and Prentice (2011), Cox and Oakes (1984)). Assume that the ith subject's response follows the following AFT model:

$$\log T_i = \mu_1 I_i + \mu_2(1-I_i) + \beta_1 X_{i,1} + \cdots + \beta_{p_1} X_{i,p_1} + \gamma_1 Z_{i,1} + \cdots + \gamma_{p_2} Z_{i,p_2} + \varepsilon_i. \tag{2.1}$$

Here μ_1 and μ_2 are parameters measuring the main effects of treatments 1 and 2; $(\beta_1, \ldots, \beta_{p_1})$ and $(\gamma_1, \ldots, \gamma_{p_2})$ are unknown parameters; the ε_is are random errors with mean zero and variance σ_ε^2. We assume that $(\epsilon_i, C_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, are independent random vectors and identically distributed as $(\epsilon, C, \boldsymbol{X}, \boldsymbol{Z})$, and, for each i, ϵ_i is independent of C_i, $X_{i,k}$ and $Z_{i,j}$, $k = 1, \ldots, p_1$, $j = 1, \ldots, p_2$, all the random variables have finite variances, and, the covariance matrix $Var(\boldsymbol{X})$ of $\boldsymbol{X}$ is nonsingular.

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, $\boldsymbol{T} = (T_1, T_2, \ldots, T_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$, $\boldsymbol{\beta} = (\mu_1, \mu_2, \beta_1, \ldots, \beta_{p_1})^T$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_2})^T$. Moreover,

$$\boldsymbol{X} = \begin{bmatrix} I_1 & 1-I_1 & X_{1,1} & \dots & X_{1,p_1} \\ I_2 & 1-I_2 & X_{2,1} & \dots & X_{2,p_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I_n & 1-I_n & X_{n,1} & \dots & X_{n,p_1} \end{bmatrix} \text{ and } \boldsymbol{Z} = \begin{bmatrix} Z_{1,1} & \dots & Z_{1,p_2} \\ \vdots & \ddots & \vdots \\ Z_{n,1} & \dots & Z_{n,p_2} \end{bmatrix}.$$

Then model (2.1) can be written as

$$\log \boldsymbol{T} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

and the working AFT model is

$$E[\log T_i] = \mu_1 I_i + \mu_2(1-I_i) + \beta_1 X_{i,1} + \cdots + \beta_{p_1} X_{i,p_1},$$

i.e.,

$$\log T_i = \mu_1 I_i + \mu_2(1-I_i) + \beta_1 X_{i,1} + \cdots + \beta_{p_1} X_{i,p_1} + u_i, \tag{2.2}$$

where $u_i = \gamma_1 Z_{i,1} + \cdots + \gamma_{p_2} Z_{i,p_2} + \varepsilon_i$ is considered as the residual. From the weighted least squares (WLS) method (Stute (1993, 1996); Shen et al. (2009)), the regression coefficients can be estimated by

$$\begin{aligned} \widehat{\boldsymbol{\beta}} =& \left\{ \sum_{i=1}^{n} \frac{\delta_i \underline{\boldsymbol{X}}_i \underline{\boldsymbol{X}}_i^T}{\widehat{G}(Y_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{\delta_i \underline{\boldsymbol{X}}_i \log Y_i}{\widehat{G}(Y_i)} \\ =& \boldsymbol{\beta} + \left\{ \sum_{i=1}^{n} \frac{\delta_i \underline{\boldsymbol{X}}_i \underline{\boldsymbol{X}}_i^T}{\widehat{G}(Y_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{\delta_i \underline{\boldsymbol{X}}_i}{\widehat{G}(Y_i)} u_i, \end{aligned} \tag{2.3}$$

where $\underline{\boldsymbol{X}}_i = \left(I_i, 1-I_i, \boldsymbol{X}_i^T\right)^T = (I_i, 1-I_i, X_{i,1}, X_{i,2} \ldots X_{i,p_1})^T$. Here $\widehat{G}(\cdot)$ is the Kaplan–Meier estimator of the survival function $G(\cdot)$ which is defined (c.f. Gill (1983) or (3.2.6) of Gill (1980))by

$$\widehat{G}(t) = \prod_{s \le t} \left(a - \frac{dN_c(s)}{Y(s)}\right),$$

$$Y(t) = \#\{Y_i \ge t\} \quad \text{and} \quad N_c(t) = \#\{Y_i \le t, \delta_i = 0\}.$$

Then the estimator of the residual u_i is $\widehat{u}_i = \log Y_i - \underline{\boldsymbol{X}}_i \widehat{\boldsymbol{\beta}}$, and the WLS estimator of the variance-covariance of $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{Var}}_{WLS}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \widehat{\Gamma}_{\boldsymbol{\beta}}^{-1} \widehat{\Sigma}_{\boldsymbol{\beta},WLS} \widehat{\Gamma}_{\boldsymbol{\beta}}^{-1},$$

where

$$\widehat{\Gamma}_{\boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \underline{\boldsymbol{X}}_i \underline{\boldsymbol{X}}_i^T}{\widehat{G}(Y_i)}, \tag{2.4}$$

$$\widehat{\Sigma}_{\boldsymbol{\beta},WLS} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i^2 \underline{\boldsymbol{X}}_i \underline{\boldsymbol{X}}_i^T}{\widehat{G}^2(Y_i)} \widehat{u}_i^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i^2 \underline{\boldsymbol{X}}_i \underline{\boldsymbol{X}}_i^T}{\widehat{G}^2(Y_i)} \left(\log T_i - \underline{\boldsymbol{X}}_i \widehat{\boldsymbol{\beta}}\right)^2. \tag{2.5}$$

In general, the WLS estimator $\widehat{\Sigma}_{\boldsymbol{\beta},WLS}$ may be biased because of the estimating the $G(\cdot)$. A valid estimator is given by

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \widehat{\Gamma}_{\boldsymbol{\beta}}^{-1} \left(\widehat{\mathrm{Var}}_{WLS}(\widehat{\boldsymbol{\beta}}) - \widehat{\Sigma}_{\boldsymbol{\beta},G}\right) \widehat{\Gamma}_{\boldsymbol{\beta}}^{-1},$$

where $\widehat{\Sigma}_{\boldsymbol{\beta},G}$ is defined by

$$\widehat{\Sigma}_{\boldsymbol{\beta},G} = \int_0^\infty \frac{\widehat{\boldsymbol{B}}_1^{\otimes 2}(s)}{\widehat{\pi}(s)} d\widehat{\Lambda}_G(s), \tag{2.6}$$

where $\widehat{\boldsymbol{B}}_1(s) = \frac{1}{n}\sum_{i=1}^n \delta_i I\{Y_i \geq s\}\underline{\boldsymbol{X}}_i(\log Y_i - \underline{\boldsymbol{X}}_i\widehat{\boldsymbol{\beta}})/\widehat{G}(Y_i)$, $\widehat{\pi}(s) = \frac{1}{n}\sum_{i=1}^n I\{Y_i \geq s\}$, and $\widehat{\Lambda}_G(s)$ is the Nelson estimate for the cumulative hazard function $\Lambda_G(s)$ of C.

We discuss clinical trials with the following hypothesis test:

$$H_0 \ : \mu_1 - \mu_2 = 0 \text{ versus } H_A \ : \mu_1 - \mu_2 \neq 0. \tag{2.7}$$

The test statistic for (2.7) is

$$\mathcal{T}(n) = \frac{\boldsymbol{L}^T\widehat{\boldsymbol{\beta}}}{\{\widehat{\mathrm{Var}}(\boldsymbol{L}^T\widehat{\boldsymbol{\beta}})\}^{1/2}}, \tag{2.8}$$

where $\boldsymbol{L} = (1, -1, 0, .., 0)^T$ and $\widehat{\mathrm{Var}}(\boldsymbol{L}^T\widehat{\boldsymbol{\beta}}) = \boldsymbol{L}^T\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\boldsymbol{L}$. We will show (see S1.22) that under the null or local alternative hypothesis,

$$\begin{aligned} n\widehat{\mathrm{Var}}(\boldsymbol{L}^T\widehat{\boldsymbol{\beta}}) &= n\boldsymbol{L}^T\widehat{\mathrm{Var}}_{WLS}(\widehat{\boldsymbol{\beta}})\boldsymbol{L} + o_P(1) \\ &= \frac{4}{n}\sum_{i=1}^n \frac{\delta_i}{\widehat{G}^2(Y_i)}\big(\log T_i - \underline{\boldsymbol{X}}_i\widehat{\boldsymbol{\beta}}\big)^2 + o_P(1). \end{aligned}$$

Hence, in (2.8) we can use $\frac{4}{n^2}\sum_{i=1}^n \frac{\delta_i}{\widehat{G}^2(Y_i)}\big(\log T_i - \underline{\boldsymbol{X}}_i\widehat{\boldsymbol{\beta}}\big)^2$ as the estimator of $\mathrm{Var}(\boldsymbol{L}^T\widehat{\boldsymbol{\beta}})$.

2.2 CAR designs

In clinical trials, CAR designs are usually based on discrete covariates (Taves (2010)). If a continuous covariate is to be used in the randomization, it need to be discretized. Define $H^* = \{k | X_k \text{ is continuous}, k = 1, \ldots, p_1\}$

and $H = \{j|Z_j \text{ is continuous}, j = 1, \ldots, p_2\}$, $d_k^*(X_k)$. Let $d_j(Z_j)$ be discrete functions, and define

$$\tilde{X}_k = \begin{cases} X_k & \text{if } k \notin H^* \\ d_k^*(X_k) & \text{if } k \in H^* \end{cases}$$

and

$$\tilde{Z}_j = \begin{cases} Z_j & \text{if } j \notin H \\ d_j(Z_j) & \text{if } j \in H. \end{cases}$$

The CAR design will assign the $(m+1)$th patient based on $\tilde{X}_{i,k}, \tilde{Z}_{i,j}, i = 1, ..., m+1$ and $I_i, i = 1, ..., m$.

Suppose $\tilde{X}_k$ has s_k^* levels and $\tilde{Z}_j$ has s_j levels. Let $W_i = \left(\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,p_1}, \tilde{Z}_{i,1}, \ldots, \tilde{Z}_{i,p_2}\right)$ represent the ith patient's covariate profile used in the CAR designs. We use $(t_1, t_2, \ldots, t_{p_1}, r_1, r_2, \ldots, r_{p_2})$ to denote the stratum formed by the patients with the same covariate levels $x_k^{t_k}$ for $\tilde{X}_k, k = 1, \ldots, p_1$ and $z_j^{r_j}$ for $\tilde{Z}_j, j = 1, \ldots, p_2$. Let $(k; t_k)$ be the margin formed by patients with level $x_k^{t_k}$ for covariate $\tilde{X}_{i,k}$ and $(j; r_j)$ be that formed by patients with level $z_j^{r_j}$ for covariate $\tilde{Z}_{i,j}$. We next introduce the measures for imbalances. In some trials, unbalanced allocations are required. For example, one may be willing to allocate more patients to the treatment group than to the placebo group. We therefore consider the q-balance, in which the number of patients in group 1 is nearly $q \times 100\%$ of the total number of patients,

$0 < q < 1$. We introduce the following notation:

1. $D_n^{(q)}$: difference between the number of patients in treatment group 1 and $q \times 100\%$ of the overall number of patients;

2. $D_n^{(q)}(X, k; t_k)$ and $D_n^{(q)}(Z, j; r_j)$: difference between the number of patients in group 1 and $q \times 100\%$ of the total number of patients on the margins $(k; t_k)$ and $(j; r_j)$, respectively;

3. $D_n^{(q)}(t_1, t_2, \ldots, t_{p_1}, r_1, r_2, \ldots, r_{p_2})$: difference between the number of patients in group 1 and $q \times 100\%$ of the total number of patients in the stratum $(t_1, t_2, \ldots, t_{p_1}, r_1, r_2, \ldots, r_{p_2})$.

When $q = 1/2$, $D_n = 2D_n^{(q)}$, $D_n(X, k; t_k) = 2D_n^{(q)}(X, k; t_k)$, and $D_n(t_1, \ldots, t_{p_1}, r_1, \ldots, r_{p_2}) = D_n^{(q)}(t_1, \ldots, t_{p_1}, r_1, \ldots, r_{p_2})$ are the usual imbalance measures when balance is required.

2.3 Main results

When CAR designs are applied in clinical trials, the main concerns are whether traditional tests are still valid with well-controlled type I error rates due to the dependence among the responses, treatment assignments, and covariates. The primary purpose of this section is to derive the asymptotic properties of hypothesis tests following CAR designs under both the

null hypothesis and the alternative hypothesis. A test is said to be (asymptotically) conservative if the true type I error is smaller than the significance level under the null hypothesis. We have the following theorem for comparing treatment effects and performing test (2.7).

Theorem 1. *Suppose that a covariate-adaptive design satisfies the following three conditions:*

(A) $Cov(X_{i,k}, u_i) = 0$, $k = 1, \ldots, p_1$;

(B) $\sum_{i=1}^{n}(I_i - q)\breve{u}_i = o_P(\sqrt{n})$, *where* $\breve{u}_i = E[u_i|W_i] - E[u_i]$;

(C) the within-stratum q-imbalances for all covariates are of order $o(n)$ *in probability, i.e.,* $D_n^{(q)}(t_1, \ldots, t_{p_1}, r_1, \ldots, r_{p_2}) = o_P(n)$ *for all* t_k*s and* r_j*s.*

Further, suppose that the regularity conditions (Ra)–(Rc) in the Supplementary Material are satisfied. Then we have the following results:

(i) Under $H_0 : \mu_1 - \mu_2 = 0$,

$$\mathcal{T}(n) \xrightarrow{D} N(0, \tau^2), \text{ with } \tau^2 = \frac{\sigma^2_{\delta,G}}{\sigma^2_{z,G}} \tag{2.9}$$

where $\sigma^2_{z,G} = E\big[(u_i - Eu_i)^2/G(T_i \wedge \tau_G)|H_0\big]$, $\sigma^2_{\delta,G} = \sigma^2_{z,G} - E\breve{u}_i^2$.

(ii) Under $H_A : \mu_1 - \mu_2 \neq 0$*, consider a sequence of local alternatives, i.e.,* $\mu_2 = \mu_1 - \delta/\sqrt{n}$ *for a fixed* $\delta \neq 0$*. Then*

$$\mathcal{T}(n) \xrightarrow{D} N(\Delta, \tau^2), \text{ with } \Delta = \frac{\delta\sqrt{q(1-q)}}{\sigma_{z,G}}. \tag{2.10}$$

Assumption (A) is used to make the parameters $\mu_1 - \mu_2$, $\beta_1, \ldots, \beta_{p_1}$ in the working model (2.2) identifiable. Assumption (B) is a general condition. The fist remark tell us that when it is satisfied under the marginal CARs or stratified CARs.

Remark 1. (i) Suppose that the marginal q-imbalances for covariates $Z_1, \ldots, Z_{p_2}$ are of order $o(\sqrt{n})$ in probability, i.e., $D_n^{(q)}(Z, j; r_j) = o_P(\sqrt{n}), j = 1, \ldots, p_2$, and that $Z_1, \ldots, Z_{p_2}$ are independent and independent of $\boldsymbol{X}$. Then Assumptions (A) and (B) are satisfied. In this case, $E[\breve{u}_i^2] = \sum_{j=1}^{p_2} \gamma_j^2 Var(E[Z_{i,j}|\tilde{Z}_{i,j}])$.

(ii) Suppose that the within-stratum q-imbalances for $Z_1, \ldots, Z_{p_2}$ are of order $o(\sqrt{n})$ in probability, i.e., $D_n^{(q)}(Z; r_1, \ldots, r_{p_2}) = o_P(\sqrt{n})$ for all r_js, and that $\boldsymbol{Z}$ is independent of $\boldsymbol{X}$. Then Assumptions (A) and (B) are satisfied.

(iii) Suppose the order $o(n)$ in Assumption (C) is strengthened to $o(\sqrt{n})$. Then Assumptions (B) and (C) are satisfied.

Further, when Assumption (A) is not satisfied, we can project u_i to the linear space spanned by $\{X_{i,1}, \cdots, X_{i,p_1}\}$ such that $u_i^* = u_i - \sum_{k=1}^{p_1} \tilde{\beta}_k X_{i,k}$

satisfies Assumption (A). Then the working model becomes

$$\log T_i = I_i\mu_1 + (1 - I_i)\mu_2 + \sum_{k=1}^{p_1} \beta_k^* X_{i,k} + u_i^*, \tag{2.11}$$

where $\beta_k^* = \beta_k + \tilde{\beta}_k$, and, Assumptions (A)-(C) are satisfied and Theorem 1 holds with μ_i^* taking the place of u_i. In particular, the test is valid in the sense of that $\lim_{n\to\infty} \mathrm{P}(|\mathcal{T}(n)| > Z_{1-\alpha/2}|H_0) \leq \alpha$, where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of a standard normal distribution.

Remark 2. Recently, for time-to-event outcomes under CARs, Ye and Shao (2020) proposed log-rank test and robust score test, Wang et al. (2023) considered Kaplan-Meier estimator, and Ye et al. (2022) proposed a covariate-adjusted log-rank test for treatment effects. These papers do not make any assumptions on the true outcome generating process and the test is model-free. When the order $o(n)$ in Assumption (C) is strengthened to $o(\sqrt{n})$, our test $\mathcal{T}(n)$ can be regarded as a model-free test for treatment effect, because under the null alternative hypothesis that the distribution of T_i is the same for $I_i = 1$ and $I_i = 0$, $\log T_i$ can be written as the form of (2.11) with $\mu_1 = \mu_2$ by projecting $\log T_i$ to the linear space spanned by $\{1, X_{i,1}, \ldots, X_{i,p_1}\}$. Theorem 1 (i) remains true under the the null hypothesis, but Theorem 1 (ii) no longer holds.

Remark 3. If $C_i = \infty$, i.e., there is no censoring, then $G(t) \equiv 1$, and

$\sigma^2_{z,G} = Var(u_i)$, $\sigma^2_{\delta,G} = Var(u_i - E[u_i|W_i])$.

Remark 4. The values of $\sigma_{z,G}$, $\sigma_{\delta,G}$, and τ do not depend on q. However, $|\Delta|$ depends on q and takes its largest value $\frac{|\delta|}{2\sigma_{z,G}}$ when $q = 1/2$. This means that the test will lose its power when $q \neq 1/2$ under the local alternatives.

Remark 5. The weighted least square method introduced Kaplan–Meier weights for the estimators, leading to significant complexity in deriving the asymptotic properties of the test statistics. We show in the Supplementary Material that the Kaplan–Meier estimator $\widehat{G}(\cdot)$ of the underlying survival function $G(\cdot)$ has asymptotically the same contribution to the estimators of μ_1 and μ_2, and the contribution is canceled in the estimator of $\mu_1 - \mu_2$ so that the Kaplan–Meier estimator has no impact on the asymptotic variance of the estimator of $\mu_1 - \mu_2$ under both the null hypothesis and the local alternative hypothesis.

Theorem 2. *Suppose that the conditions of Theorem 1 are satisfied. Then, if the test statistic* (2.8) *is used to perform the hypothesis test* (2.7)*, we have the following results:*

(1) A valid type I error rate can be obtained if all the randomization covariates are included in the data analysis.

(2) The type I error rate is conservative if not all the randomization co-

variates are included in the analysis and $E[u_i|W_i] \not\equiv Constant$*. That is, for a given significance level* α*, there is a constant* α_0 *such that, when* H_0 *holds,* $\lim_{n\to\infty} P(|\mathcal{T}(n)| > Z_{1-\alpha/2}) \leq \alpha_0 < \alpha$.

This echoes Forsythe's recommendation (Forsythe, 1987) that in time-to-event data analysis all variables used in the minimization should also be used as covariates in the analysis.

From Theorems 1 and 2, the balance of the covariates plays an important role in inference for covariate-adaptive designs. For the stratified permuted block design, the difference within any stratum is at most half of the block size. Since the number of strata is finite, the overall and marginal differences are less than a constant, so conditions (A), (B), and (C) are satisfied. Hu and Hu (2012) proposed a large class of covariate-adaptive designs in which the overall difference, marginal differences, and within-stratum imbalance are all bounded in probability, i.e., $D_n = O_P(1)$, $D_n(X, k; t_k) = O_P(1)$, $D_n(Z, j; r_j) = o_P(1)$, and $D_n(t_1, \ldots, t_{p_1}, r_1, \ldots, r_{p_2}) = O_P(1)$, under certain conditions, and conditions (A), (B), and (C) are satisfied. For Pocock and Simon's marginal procedure (Pocock and Simon, 1975), the marginal difference and the overall difference have been proved by Ma et al. (2015) to be bounded in probability, and the within-stratum imbalance has been proved by Hu and Zhang (2020) to be of order $O(\sqrt{n})$ in probability.

We summarize these results in the corollary below.

Corollary 1. *Suppose $q = 1/2$ and that the regularity conditions (Ra) and (Rb) in the Supplementary Material are satisfied. Theorems 1 and 2 hold under the following covariate-adaptive designs:*

(i) *stratified permuted block designs;*

(ii) *the class of covariate-adaptive designs proposed by Hu and Hu (2012);*

(iii) *Pocock and Simon's marginal procedure (Pocock and Simon, 1975) with the Assumption that $Z_1, \ldots, Z_{p_2}$ are independent and independent of $\boldsymbol{X}$.*

In addition to the above conclusions concerning inference under our procedure, we offer more details about the consistency and asymptotic normality of the parameter estimators.

Theorem 3. *Under assumption (C) of Theorem 1 and regularity conditions (Ra)–(Rc), $\widehat{\boldsymbol{\beta}}$ is a consistent estimate of $\boldsymbol{\beta}$.*

Theorem 4. *Under the conditions of Theorem 1,*

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) \xrightarrow{D} N\big(0, \Gamma_{\boldsymbol{\beta}}^{-1}\Sigma_{\boldsymbol{\beta}}\Gamma_{\boldsymbol{\beta}}^{-1}\big), \tag{2.12}$$

where $\Gamma_{\boldsymbol{\beta}}$ is the limit of $\widehat{\Gamma}_{\boldsymbol{\beta}}$ and is defined as in (2.4), $\Sigma_{\boldsymbol{\beta}}$ is defined as in (S1.14) and $\Sigma_{\boldsymbol{\beta}}+q(1-q)\boldsymbol{L}\boldsymbol{L}^T E(\breve{u}_i)^2$ is the limit of $\widehat{\Sigma}_{\boldsymbol{\beta}} = \widehat{Var}_{WLS}(\widehat{\boldsymbol{\beta}}) - \widehat{\Sigma}_{\boldsymbol{\beta},G}$.

We also consider general forms of hypothesis testing for the significance of the covariates. Let $\mathcal{P}$ be an $m \times (p_1+2)$ matrix of rank m with $m < (p_1+2)$, where the entries of the first two columns are all zero. Our hypothesis would be

$$H_0' : \mathcal{P}\boldsymbol{\beta} = \boldsymbol{\xi}_0 \text{ versus } H_A' : \mathcal{P}\boldsymbol{\beta} = \boldsymbol{\xi}_1 \neq \boldsymbol{\xi}_0. \tag{2.13}$$

The test statistic for (2.13) is

$$\mathcal{T}_\beta = (\mathcal{P}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}_0)^T [\mathcal{P}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathcal{P}^T]^{-1}(\mathcal{P}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}_0). \tag{2.14}$$

It can be shown that $\mathcal{P}\Gamma_{\boldsymbol{\beta}}^{-1}\boldsymbol{L} = \mathbf{0}$. Then

$$\begin{aligned}\lim_{n\to\infty} \mathcal{P}\big(n\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\big)\mathcal{P}^T &= \lim_{n\to\infty} \mathcal{P}\Gamma_{\boldsymbol{\beta}}^{-1}\Big(\Sigma_{\boldsymbol{\beta}} + q(1-q)\boldsymbol{L}\boldsymbol{L}^T E(\breve{u}_i)^2\Big)\Gamma_{\boldsymbol{\beta}}^{-1}\mathcal{P}^T \\ &= \mathcal{P}\Gamma_{\boldsymbol{\beta}}^{-1}\Sigma_{\boldsymbol{\beta}}\Gamma_{\boldsymbol{\beta}}^{-1}\mathcal{P}^T,\end{aligned}$$

which is the asymptotic covariance matrix of $\sqrt{n}\mathcal{P}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)$. Hence have the following theorem.

Theorem 5. *Under the conditions of Theorem 1, we have the following results:*

(i) Under $H_0' : \mathcal{P}\boldsymbol{\beta} = \boldsymbol{\xi}_0$,

$$\mathcal{T}_\beta \xrightarrow{D} \chi^2_{(m)}.$$

Hence, the hypothesis test for (2.13) *can achieve the correct Type I error.*

(ii) Under $H_A' : \mathcal{P}\boldsymbol{\beta} = \boldsymbol{\xi}_1$, *consider a sequence of local alternatives, i.e.,* $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_0 + \boldsymbol{\eta}/\sqrt{n}$ *for a fixed* $\boldsymbol{\eta} \neq \mathbf{0}$. *Then*

$$\mathcal{T}_\beta \xrightarrow{D} \chi^2_{(m)}(\lambda), \quad \textit{with} \quad \lambda = \boldsymbol{\eta}^T[\mathcal{P}\boldsymbol{M}^{-1}\mathcal{P}^T]^{-1}\boldsymbol{\eta},$$

where $\boldsymbol{M} = \lim_{n\to\infty} n\widehat{Var}(\widehat{\boldsymbol{\beta}})$ *and* λ *is the noncentral parameter.*

3. Numerical studies

In this section, we discuss how to control the type I error rate and investigate the finite-sample performance of our methods. Herein, we use $\mathcal{T}(n)/\hat{\tau}$ as the test statistic, where $\hat{\tau}$ is a consistent estimator of τ. We fit model (2.1) with full data to obtain consistent estimators of all the unknown parameters. We compare Pocock and Simon's design (PS), stratified permuted block design (SPB), and complete randomization (CR); we use either the ordinary test statistic (2.8) or our adjusted test statistic.

We assume that the survival time T_i follows the following AFT model:

$$\log(T_i) = \mu_1(1 - I_i) + \mu_2 I_i + \beta_1 Z_{i,1} + \beta_2 Z_{i,2} + \epsilon_i, \tag{3.15}$$

where $\beta_1 = \beta_2 = 0.5$, $Z_{i,1}$ and $Z_{i,2}$ are independent. For simplicity, we do not distinguish the notation of $\boldsymbol{X}$ and $\boldsymbol{Z}$. The following two models are fitted to the data: (1) no covariates are included in the AFT model

(AFT), and (2) both Z_1 and Z_2 are included ($AFT(Z_1, Z_2)$). We report the parameter estimates ($\hat{\mu}_2$ as a representative) and the type I error rate. We also compare the estimated standard deviation (Est.Sd) and the empirical standard deviation (Emp.Sd) of the estimated treatment effects $\hat{\mu}_1 - \hat{\mu}_2$.

In Table 1, both covariates are binary with a success rate of 0.5, and ϵ_i follows $N(0, 0.25^2)$. The censoring time is generated from a uniform distribution on $(0, c)$, and $c = 10$ is chosen to make the censoring rate approximately 20%. We can see that if we include all the randomization covariates in the analysis, the type I error rate can be controlled, and the parameter can be accurately estimated for both complete randomization and CAR designs. The Est.Sd of the estimated treatment effects without any adjustment as in (2.8) and the Emp.Sd are similar. However, if we do not use the randomization covariates in the AFT model, the type I error rates are conservative under CAR designs, and the Emp.Sd of the estimated treatment effects is smaller than the Est.Sd. In other words, CAR designs reduce the variability of the estimated treatment effect. When using our adjusted test statistics, we can control the type I error rate well for CAR designs. Moreover, the Est.Sd after our adjustment is quite consistent with the Emp.Sd, which explains why our methods can control the type I error rate well. We also found that our methods can accurately estimate the

Table 1: Performance of our methods when error follows normal distribution and both covariates are binary

(μ_1, μ_2)	Model	Allocation	Type I error	Est.Sd	Emp.Sd	$\hat{\mu}_2$
$(0.5, 0.5)$	AFT(Z_1, Z_2)	CR	0.050	0.025	0.025	0.500
		PS	0.054	0.025	0.025	0.500
		SPB	0.051	0.025	0.025	0.500
	AFT	CR	0.051	0.044	0.045	0.500
		PS	0.007	0.044	0.031	0.500
		PS_{adj}	0.054	0.031	0.031	0.500
		SPB	0.006	0.044	0.031	0.499
		SPB_{adj}	0.054	0.031	0.031	0.499
$(0.4, 0.4)$	AFT(Z_1, Z_2)	CR	0.051	0.025	0.025	0.400
		PS	0.054	0.025	0.025	0.400
		SPB	0.054	0.025	0.025	0.400
	AFT	CR	0.056	0.043	0.044	0.399
		PS	0.007	0.043	0.030	0.400
		PS_{adj}	0.054	0.030	0.030	0.400
		SPB	0.004	0.043	0.030	0.400
		SPB_{adj}	0.053	0.030	0.030	0.400

unknown parameter.

In Table 2, we study the performance of our methods when the errors

follow the logistic distribution with location parameter 0 and scale parameter α. In Table 3, we study the case where both covariates follow the Normal distribution $N(0, 0.5^2)$. The censoring time is generated from a uniform distribution on $(0, 8)$. All the other settings are as in Table 1. Our conclusions are similar to those for the binary covariates (Table 1).

In Table 4, we study the performance of our methods in terms of the hypothesis test for the covariates under $H_0 : \beta_1 = 0$ when both covariates are binary with success rate 0.5 and ϵ_i follows $N(0, 0.25^2)$. The censoring time is generated from a uniform distribution on $(0, 8)$. The following two models will be fitted to the data: (1) only Z_1 is included in the AFT model ($AFT(Z_1)$), and (2) both Z_1 and Z_2 are included ($AFT(Z_1, Z_2)$). All the other settings are as in Table 1. Our theorem shows that we can control the type I error rate without adjustment, and Table 4 confirms these theoretical results. For simplicity, we report only the type I error rate here.

We have carried out numerical studies for various combinations of the distributions of the errors and covariates, the values of the unknown parameters, and hypothesis tests for the treatment or covariate effects. All led to similar conclusions, and the details are omitted.

Table 2: Performance of our methods when error follows logistic distribution and both covariates are binary

(μ_1, μ_2, α)	Model	Allocation	Type I error	Est.Sd	Emp.Sd	$\hat{\mu}_2$
$(0.4, 0.4, 0.12)$	AFT(Z_1, Z_2)	CR	0.052	0.021	0.022	0.400
		PS	0.047	0.021	0.021	0.399
		SPB	0.049	0.021	0.022	0.399
	AFT	CR	0.050	0.041	0.041	0.399
		PS	0.003	0.041	0.027	0.399
		PS_{adj}	0.052	0.027	0.027	0.399
		SPB	0.002	0.041	0.027	0.399
		SPB_{adj}	0.052	0.027	0.027	0.399
$(0.6, 0.6, 0.15)$	AFT(Z_1, Z_2)	CR	0.048	0.028	0.028	0.599
		PS	0.054	0.028	0.028	0.599
		SPB	0.054	0.028	0.028	0.599
	AFT	CR	0.049	0.046	0.046	0.599
		PS	0.007	0.046	0.034	0.599
		PS_{adj}	0.052	0.034	0.034	0.599
		SPB	0.009	0.046	0.034	0.599
		SPB_{adj}	0.050	0.034	0.034	0.599

Table 3: Performance of our methods when error and both covariates follow normal distribution

(μ_1, μ_2)	Model	Allocation	Type I error	Est.Sd	Emp.Sd	$\hat{\mu}_2$
$(0.4, 0.4)$	AFT(Z_1, Z_2)	CR	0.050	0.025	0.025	0.400
		PS	0.050	0.025	0.026	0.400
		SPB	0.050	0.025	0.026	0.399
	AFT	CR	0.052	0.045	0.046	0.399
		PS	0.022	0.045	0.038	0.399
		PS_{adj}	0.053	0.038	0.038	0.399
		SPB	0.019	0.045	0.038	0.399
		SPB_{adj}	0.049	0.038	0.038	0.399
$(0.3, 0.3)$	AFT(Z_1, Z_2)	CR	0.055	0.025	0.025	0.300
		PS	0.052	0.025	0.025	0.300
		SPB	0.051	0.025	0.025	0.299
	AFT	CR	0.053	0.045	0.045	0.299
		PS	0.019	0.044	0.037	0.299
		PS_{adj}	0.050	0.037	0.037	0.299
		SPB	0.024	0.045	0.038	0.299
		SPB_{adj}	0.055	0.037	0.038	0.299

Table 4: Performance of our methods for hypothesis tests about covariate effects when error follows normal distribution and both covariates are binary

(μ_1, μ_2, β_2)	Allocation	AFT(Z_1, Z_2)	AFT(Z_1)
$(0.5, 0.5, 0.5)$	CR	0.053	0.050
	PS	0.052	0.046
	SPB	0.046	0.050
$(0.5, 0.6, 0.7)$	CR	0.047	0.050
	PS	0.048	0.051
	SPB	0.054	0.047

4. Conclusion

CAR designs, especially stratified permuted block randomization designs, are popular in clinical trials since balancing the treatment allocation for influential covariates is important. Two questions about CAR designs are of interest. First, what are the asymptotic imbalances at different levels: within strata, marginal, and overall? Hu and Hu (2012) used the Markov technique to derive the order of these imbalances for a family of CAR designs. Ma et al. (2015) proved the asymptotic imbalance for both Pocock and Simon's marginal procedures (Pocock and Simon, 1975) and stratified

permuted block designs.

Second, why is the type I error rate conservative when not all the covariates are included in the analysis? This has been numerically demonstrated in different settings. However, there has been little research into the performance of hypothesis testing for trials with time-to-event outcomes under a general family of CAR designs. We have provided the theoretical properties of hypothesis testing under the AFT model with CAR designs. We derived the asymptotic distribution of the test statistic under both the null and alternative hypotheses and showed the consistency and asymptotic normality of the estimators. We explained the conservativeness of the type I error rate when only some of the randomization covariates are used in the model. Based on our theoretical results, we proposed methods to control the type I error rate. Numerical results confirmed our theoretical findings and demonstrated the success of our methods.

This paper opens the door to further study of clinical trials with CAR designs and time-to-event outcomes. There are several future research directions. First, modern trials often require an interim analysis, and we need the joint distribution of the sequential statistics. The current paper offers the marginal distribution of each of these sequential statistics. Second, both the industry and the FDA demand the evaluation of new therapies in

a time-sensitive and cost-effective manner. We can study adaptive seamless phase II/III clinical trial designs (ASD) with CAR and AFT to satisfy this need for time-to-event responses. Based on different approaches for ASD, we may need to generalize our results to trials with multiple treatments and investigate the correlation of statistics from different phases. Third, robust inference under CAR designs has recently attracted much attention, and it would be interesting to explore robust inference for clinical trials with CAR and survival responses.

Acknowledgments The authors thank the Associate Editor and the two anonymous referees for many helpful comments and suggestions. This publication was neither originated nor managed by AbbVie, and it does not communicate results of AbbVie-sponsored Scientific Research. Thus, it is not in scope of the AbbVie Publication Procedure (PUB-100).

References

Aickin, M. (2002). Beyond randomization. *The Journal of Alternative & Complementary Medicine 8*(6), 765–772.

Antognini, A. B. and M. Zagoraiou (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika 98*(3), 519–535.

Ashley, E. A., A. J. Butte, M. T. Wheeler, R. Chen, T. E. Klein, F. E. Dewey, J. T. Dudley,

K. E. Ormond, A. Pavlovic, A. A. Morgan, et al. (2010). Clinical assessment incorporating a personal genome. *The Lancet 375*(9725), 1525–1535.

Birkett, N. J. (1985). Adaptive allocation in randomized controlled trials. *Controlled Clinical Trials 6*(2), 146–155.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*(3), 429–436.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1784–1796.

Cheng, S., L. J. Wei, and Z. Ying (1995). Analysis of transformation models with censored data. *Biometrika 82*(4), 845–845.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. CRC Press.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics 4*, 3895–3962.

Forsythe, A. B. (1987). Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics & Data Analysis 5*(3), 193–200.

Gill, R. D. (1980). *Censoring and stochastic integrals,* Mathematical Centre Tracts, Volume

124. Mathematisch Centrum, Amsterdam.

Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *The Annals of Statistics 11*(1), 49–58.

Hagino, A., C. Hamada, I. Yoshimura, Y. Ohashi, J. Sakamoto, and H. Nakazato (2004). Statistical comparison of random allocation methods in cancer clinical trials. *Controlled Clinical Trials 25*(6), 572–584.

Heritier, S., V. Gebski, and A. Pillai (2005). Dynamic balancing randomization in controlled clinical trials. *Statistics in Medicine 24*(24), 3729–3741.

Hu, F. and L.-X. Zhang (2020). On the theory of covariate-adaptive designs. *arXiv preprint arXiv:2004.02994*.

Hu, Y. and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics 40*(3), 1794–1815.

Jin, Z., D. Lin, L. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. *Biometrika 90*(2), 341–353.

Kahan, B. C., V. Jairath, C. J. Doré, and T. P. Morris (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials 15*(1), 139.

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

Koul, H., V. Susarla, and J. Van Ryzin (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics 9*, 1276–1288.

Lai, T. L. and Z. Ying (1991). Large sample theory of a modified buckley–james estimator for regression analysis with censored data. *The Annals of Statistics 19*, 1370–1402.

Lal, B. K., K. W. Beach, G. S. Roubin, H. L. Lutsep, W. S. Moore, M. B. Malas, D. Chiu, N. R. Gonzales, J. L. Burke, M. Rinaldi, et al. (2012). Restenosis after carotid artery stenting and endarterectomy: A secondary analysis of crest, a randomised controlled trial. *The Lancet Neurology 11*(9), 755–763.

Lebowitsch, J., Y. Ge, B. Young, and F. Hu (2012). Generalized multidimensional dynamic allocation method. *Statistics in Medicine 31*(28), 3537–3544.

Leon, L. F., T. Cai, and L. Wei (2009). Robust inferences for covariate effects on survival time with censored linear regression models. *Statistics in Biosciences 1*(1), 50–64.

Lin, D. and Z. Ying (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal of Statistical Planning and Inference 44*(1), 47–63.

Ma, W., F. Hu, and L. Zhang (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association 110*(510), 669–680.

Ma, W., Y. Qin, Y. Li, and F. Hu (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association 115*(531), 1488–1497.

Miller, R. and J. Halpern (1982). Regression with censored data. *Biometrika 69*(3), 521–531.

Nordle, O. and B. Brantmark (1977). A self-adjusting randomization plan for allocation of patients into two treatment groups. *Clinical Pharmacology and Therapeutics 22*(6), 825–830.

Pocock, S. J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics 31*, 103–115.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics 18*, 303–328.

Rosenberger, W. F. and J. M. Lachin (2015). *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons.

Rosenberger, W. F. and O. Sverdlov (2008). Handling covariates in the design of clinical trials. *Statistical Science 23*, 404–419.

Russell, D., Z. Hoare, R. Whitaker, C. Whitaker, and I. Russell (2011). Generalized method for adaptive randomization in clinical trials. *Statistics in Medicine 30*(9), 922–934.

Shao, J. and X. Yu (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics 69*(4), 960–969.

Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika 97*(2), 347–360.

Shen, Y., J. Ning, and J. Qin (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Asso-*

ciation 104(487), 1192–1202.

Signorini, D., O. Leung, R. Simes, E. Beller, V. Gebski, and T. Callaghan (1993). Dynamic balanced randomization for clinical trials. *Statistics in Medicine 12*(24), 2343–2350.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis 45*(1), 89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics 23*, 461–471.

Sverdlov, O. (2015). *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*. CRC Press.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics 15*(5), 443.

Taves, D. R. (2010). The use of minimization in clinical trials. *Contemporary clinical trials 31*(2), 180–184.

Toorawa, R., M. Adena, M. Donovan, S. Jones, and J. Conlon (2009). Use of simulation to compare the performance of minimization with stratified blocked randomization. *Pharmaceutical Statistics 8*(4), 264–278.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics 18*, 354–372.

Wang, B., R. Susukida, R. Mojtabai, M. Amin-Esmaeili, and M. Rosenblum (2023). Model-

robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association 118*(542), 1152C1163.

Wei, L. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association 73*(363), 559–563.

Weir, C. J. and K. R. Lees (2003). Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in Medicine 22*(5), 705–726.

Ye, T. and J. Shao (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *Journal of the Royal Statistical Society: Series B (Statistical Methodol- ogy 82*(5), 1301–1323.

Ye, T., J. Shao, and Y. Yi (2022). Covariate-adjusted log-rank test: guaranteed efficiency gain and universal applicability. *arXiv preprint arXiv:2201.11948*.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics 21*, 76–99.