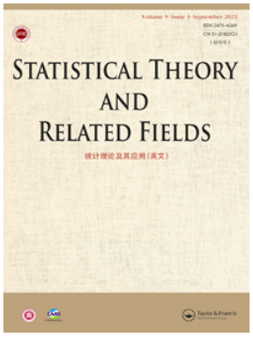

Statistical Theory and Related Fields

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

Inference after covariate-adaptive randomisation: aspects of methodology and theory

Jun Shao

To cite this article: Jun Shao (2021) Inference after covariate-adaptive randomisation: aspects of methodology and theory, Statistical Theory and Related Fields, 5:3, 172-186, DOI: 10.1080/24754269.2021.1871873

To link to this article: https://doi.org/10.1080/24754269.2021.1871873

Published online: 18 Jan 2021.

STATISTICAL THEORY AND RELATED FIELDS
2021, VOL. 5, NO. 3, 172–186
https://doi.org/10.1080/24754269.2021.1871873

Taylor & Francis
Taylor & Francis Group

Inference after covariate-adaptive randomisation: aspects of methodology and theory

Jun Shao[a,b]

[a]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Covariate-adaptive randomisation has a more than 45 years of history of applications in clinical trials, in order to balance treatment assignments across prognostic factors that may have influence on the outcomes of interest. However, almost no theory had been developed for covariate-adaptive randomisation until a paper on the theory of testing hypotheses published in 2010. In this article, we review aspects of methodology and theory developed in the last decade for statistical inference under covariate-adaptive randomisation. We focus on issues such as whether a conventional procedure valid under the assumption that treatments are assigned completely at random is still valid or conservative when the actual randomisation is covariate-adaptive, how a valid inference procedure can be obtained by modifying a conventional method or directly constructed by stratifying the covariates used in randomisation, whether inference procedures have different properties when covariate-adaptive randomisation schemes have different degrees of balancing assignments, and how to further adjust covariates in the inference procedures to gain more efficiency. Recommendations are made during the review and further research problems are discussed.

ARTICLE HISTORY
Received 10 August 2020
Revised 19 December 2020
Accepted 1 January 2021

KEYWORDS
balancedness of assignments; efficiency; model-assisted approach; model free inference; stratification; survival analysis

1. Introduction

In a clinical trial to compare $k \geq 2$ treatments, patients are typically randomised into treatment arms according to fixed treatment assignment proportions $\pi_1, \ldots, \pi_k$, where each π_t is a known number strictly between 0 and 1 and $\sum_{t=1}^{k} \pi_t = 1$. The simplest randomisation scheme assigns patients to treatments completely at random and, thus, is called complete randomisation or simple randomisation. However, simple randomisation may yield imbalance assignments, i.e. sample sizes not following the assignment proportions across some prognostic factors or covariates, e.g., institution, disease stage, prior treatment, gender and age, which are thought to have significant influence on the outcomes or responses of interest. For instance, a trial exhibiting a substantial imbalance in patient age or disease stage between two treatment arms may not pass a regulatory review even though a statistically significant treatment effect has been shown. The issue is more serious when patients are not all available for simultaneous assignment of treatments but rather arrive sequentially and must be treated immediately.

This leads to the development of covariate-adaptive randomisation (which is also referred to as dynamic allocation), i.e., treatment assignment of the ith patient is made dependent on the observed covariate value of this patient and the assignments and covariate values of all $i-1$ previously assigned patients. It should be emphasised that covariate-adaptive randomisation does not use any outcomes or responses from the $i-1$ previous patients when the ith patient is randomised to a treatment arm. Adaptive randomisation methods using outcomes or responses are not our focus and can be found, for example, in Hu and Rosenberger (2006), Zhang et al. (2007), Hu et al. (2009), Rosenberger and Lachin (2015), and the references therein. The oldest covariate-adaptive randomisation scheme is the minimisation proposed by Taves (1974) and its extensions in Pocock and Simon (1975). Other popular covariate-adaptive randomisation methods include the stratified permuted block randomisation method (Zelen, 1974), the stratified urn design (Wei, 1977; Zhao & Ramakrishnan, 2016) and the stratified biased coin method (Kuznetsova & Johnson, 2017; Shao et al., 2010). Summaries of different allocation schemes are given by Kalish and Begg (1985), Schulz and Grimes (2002), and Rosenberger and Sverdlov (2008).

How often are covariate-adaptive schemes applied in clinical trials? According to Taves (2010), from 1989 to 2008, over 500 clinical trials implemented the minimisation method to balance important covariates. In a recent review of nearly 300 clinical trials published in year 2009 and year 2014 (Ciolino et al., 2019), 237 of them used covariate-adaptive randomisation. Other

CONTACT Jun Shao shao@stat.wisc.edu

examples can be found in van der Ploeg et al. (2010), Fakhry et al. (2015), Breugom et al. (2015), Stott et al. (2017), and Sun et al. (2018). In the 2018 *New England Journal of Medicine*, there are seven articles about covariate-adaptive randomisation (Horn et al., 2018; Jourdain et al., 2018; McKeever et al., 2018; Mehra et al., 2018; Myles et al., 2018; Ramirez et al., 2018; Zannad et al., 2018). Applications of covariate-adaptive randomisation are not limited to clinical trials, as they are relevant for randomised experiments with many interventions.

How should inference be carried out with data collected under covariate-adaptive randomisation? Unfortunately, tests and other inference procedures constructed based on simple randomisation, which will be called conventional tests and inference procedures, are often applied in practice after data are collected under covariate-adaptive randomisation. For example, in the seven articles cited previously in the 2018 *New England Journal of Medicine*, they all used conventional tests for treatment effect. On one hand, over 35 years between 1974 and 2009, there were many empirical results showing that some conventional tests could still control Type I errors in spite of using covariate-adaptive randomisation; see, for example Birkett (1985), Forsythe (1987), Aickin (2002), Weir and Lees (2003), Hagino et al. (2004), and Zhong and Kim (2008). On the other hand, the Committee for Proprietary Medicinal Product commended that 'it remains controversial whether the analysis adequately reflects the randomisation scheme' (Committee for proprietary medicinal products, 2004) and the European Medicines Agency 2015 guidelines stated that 'possible implications of dynamic allocation methods [minimisation] on the analysis, e.g., with regard to bias and Type I error control should be carefully considered,… conventional statistical methods do not always control the Type I error' (EMA, 2015). Because a statistical inference procedure on treatment effects should be valid under the particular randomisation scheme used in data collection, the application of conventional inference procedures after covariate-adaptive randomisation has definitely raised concerns and controversies.

Why don't we always apply an inference procedure valid under a given covariate-adaptive randomisation scheme? In their review of covariate-adaptive randomi sation, Rosenberger and Sverdlov (2008) stated:

> Very little theoretical work has been done in this area, despite the proliferation of papers. The original source papers are fairly uninformative about theoretical properties of the procedures.

That is, the lack of theoretical work in developing valid inference procedures associated with covariate-adaptive randomisation schemes is probably the main reason why conventional procedures are applied in applications. Why is there so little theoretical work in this problem prior to 2008? Unlike simple randomisation, covariate-adaptive randomisation generates some dependence among treatment assignments, covariates and outcomes under which asymptotic distributions of treatment effect estimators (such as the difference of sample averages) could not be easily derived. Shao et al. (2010) initiated theoretical studies on the validity of statistical tests under covariate-adaptive randomisation. The following three issues are addressed in their paper:

(A) Can we develop a test procedure valid under covariate-adaptive randomisation?
(B) If we use covariate-adaptive randomisation and a conventional test procedure valid under simple randomisation, will the Type I error of the test be inflated?
(C) Is a test under covariate-adaptive randomisation more powerful than it is under simple randomisa tion?

If we have affirmative answers to Questions (A)–(C), or at least Questions (A)–(B), then the concerns and controversies about using covariate-adaptive randomisation will be largely eliminated. As the first piece of theoretical work, the results in Shao et al. (2010) are limited to certain types of tests, randomisation schemes, and models between covariates and responses. Fortunately, significant progresses in the theory of this area have been made in the last decade, e.g., Hu and Hu (2012), Shao and Yu (2013), Ma et al. (2015), Bugni et al. (2018, 2019), Ye (2018), Ma et al. (2020), Ye and Shao (2020), and Ye et al. (2020). Another stream of results is based on permutation or re-randomisation methods, e.g., Simon and Simon (2011), Kaiser (2012), and Bugni et al. (2018).

The purpose of this article is to review aspects of methodology and theory for statistical inference after covariate-adaptive randomisation. We concentrate on Questions (A)–(C) previously stated and the main results in the last decade, some of which are very recent. It is our hope that this review will provide some guidance for clinical trialists about which valid inference procedures to use for various situations, and will shed light on further research and development in this important area.

2. Covariates, outcomes and treatment effects

First, let's describe covariates and outcomes or responses under a clinical trial. Consider a clinical trial with a total of n patients that are assigned to $k \geq 2$ treatment arms denoted by $a = 1, \ldots, k$. From patient $i \in \{1, \ldots, n\}$, let $\boldsymbol{X}_i$ be the vector of all observed covariates and let $Y_i^{(a)}$ be the potential outcome or response of interest under treatment assignment $a = 1, \ldots, k$. $Y_i^{(a)}$ is called potential outcome because only one of $Y_i^{(1)}, \ldots, Y_i^{(k)}$ will be observed from patient i, as each patient receives only one treatment. Thus what we

observe from patient i is $Y_i = Y_i^{(a)}$ if treatment a is assigned to patient i.

The outcome or response $Y_i^{(a)}$ could be continuous or discrete, or a survival time. In a survival trial, censoring is typically involved so that, for patient i, $Y_i^{(a)} = \min(T_i^{(a)}, C_i^{(a)})$ together with an indicator of the event $T_i^{(a)} \le C_i^{(a)}$ are observed, where $T_i^{(a)}$ is the potential survival or failure time and $C_i^{(a)}$ is the potential censoring time, under treatment $a = 1, \ldots, k$.

Throughout, we assume the following minimal condition on covariates and outcomes.

(C1) $(Y_i^{(1)}, \ldots, Y_i^{(k)}, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, are independent and identically distributed.

Note that there is no assumption on the relationship between the covariates and potential outcomes. We allow arbitrary treatment effect heterogeneity, i.e., the effect of treatment and covariate interaction on potential outcomes.

For convenience, we use $Y^{(1)}, \ldots, Y^{(k)}, \boldsymbol{X}$ to denote the variables from a generic patient. Under (C1), $(Y^{(1)}, \ldots, Y^{(k)}, \boldsymbol{X}) \sim (Y_i^{(1)}, \ldots, Y_i^{(k)}, \boldsymbol{X}_i)$ for every i, where $\mathscr{X} \sim \mathscr{Y}$ means that $\mathscr{X}$ has the same distribution as $\mathscr{Y}$.

To assess treatment effect, we may be interested in the average treatment effect between any fixed pair of treatment arms, a and b, defined as $E(Y^{(a)} - Y^{(b)})$, where E is the population expectation and $E(Y^{(a)})$ is assumed to be well defined. Another important measure in comparing treatments a and b is the quantile treatment effect defined as $q_\tau^{(a)} - q_\tau^{(b)}$ (Firpo, 2007; Zhang et al., 2020), where $q_\tau^{(a)}$ is the τth quantile of the distribution of $Y^{(a)}$ under treatment a and τ is a fixed fraction. Quantile treatment effect is more appropriate when potential outcomes are highly skewed and is more relevant and informative than the average treatment effect when some distributional impacts have to be assessed.

Both average treatment effect and quantile treatment effect are some characteristics of the distributions of potential outcomes. In some applications, we would like to assess the treatment effect on the entire distribution of a potential outcome or the entire conditional distribution of a potential outcome given covariates. For example, in a survival analysis we may be interested in testing whether the conditional distributions of $T^{(a)}$ given $\boldsymbol{X}$ are the same for different a's.

Here we would like to make it clear that treatments may have effect not only on the marginal distributions of potential outcomes, but also the conditional distributions of $Y^{(a)}$ given $\boldsymbol{X}$, although typically treatments may not have any effect on the marginal distribution of the covariate $\boldsymbol{X}$.

So far we have not yet discussed the treatment assignment of patients. Suppose that treatment assignments are made according to some probabilistic mechanism. For patient i, let A_i be the treatment assignment indicator vector, i.e., $A_i = e_a$ if patient i is assigned to treatment a, where e_a is a vector whose ath component is 1 and rest components are 0's, $a = 1, \ldots, k$. The observed outcome from patient i is $Y_i = Y_i^{(a)}$ if and only if $A_i = e_a$, $a = 1, \ldots, k$, $i = 1, \ldots, n$.

3. Covariate-adaptive randomisation schemes

We now introduce details about how A_i's are generated according to a randomisation scheme, using or without using covariates $\boldsymbol{X}_i$'s.

Under simple randomisation, A_i's are independent of $(Y_i^{(1)}, \ldots, Y_i^{(k)}, \boldsymbol{X}_i)$'s and, further, A_i's are independent and identically distributed with $P(A_i = e_a) = \pi_a$, where P denotes the probability under a given randomness mechanism. It should be emphasised that the independence between A_i's and $(Y_i^{(1)}, \ldots, Y_i^{(k)}, \boldsymbol{X}_i)$'s means that the treatment assignments are independent of potential outcomes and covariates, not that treatments have no effect on potential outcomes or conditional distribution of $Y^{(a)}$ given $\boldsymbol{X}$ as discussed in Section 2.

Let $\boldsymbol{Z}$ be a vector of discrete covariates with finitely many levels to be utilised in covariate-adaptive randomisation. Typically, components of $\boldsymbol{Z}$ are some discrete components of $\boldsymbol{X}$ and/or some discretised continuous components of $\boldsymbol{X}$ that are thought to have significant influence on the potential outcomes. In the following, we describe some popular covariate-adaptive randomisation schemes for enforcing assignment allocation across at levels of $\boldsymbol{Z}$. In a typical covariate-adaptive randomisation scheme, for the ith patient arrived with observed $\boldsymbol{Z}_i$, the treatment assignment indicator A_i is generated depending on not only the value of $\boldsymbol{Z}_i$ but also the $\boldsymbol{Z}$-values and assignments of the previous $i-1$ patients, $i = 1, \ldots, n$.

The stratified permuted block randomisation method (Zelen, 1974) randomly assigns a block of size B patients into k arms each having $B\pi_a$ patients for every B sequentially arrived patients with $\boldsymbol{Z} = \boldsymbol{z}$, a particular level of $\boldsymbol{Z}$, where B is appropriately chosen so that $B\pi_a$'s are integers and the last block is allowed to be incomplete. This method is called stratified permuted block randomisation since randomisation is carried out within each stratum (joint level of $\boldsymbol{Z}$) to achieve balancedness of assignments across strata.

When $k = 2$ and $\pi_1 = \pi_2 = 1/2$, the stratified biased coin method (Shao et al., 2010) assigns patient i with $\boldsymbol{Z}_i = \boldsymbol{z}$ according to the biased coin randomisation

in Efron (1971),

$$P(A_i = e_1) = \begin{cases} p, & D_{i-1}(\boldsymbol{z}) < 0, \\ 1/2, & D_{i-1}(\boldsymbol{z}) = 0, \\ 1-p, & D_{i-1}(\boldsymbol{z}) > 0, \end{cases}$$

where p is a fixed constant satisfying $1/2 < p < 1$ and $D_{i-1}(\boldsymbol{z})$ is one half of the within $\boldsymbol{z}$ stratum difference between the numbers of patients in treatment 1 and treatment 2 after $i-1$ assignments have been made. An extension of the stratified biased coin for general case of $k \geq 3$ can be found in Kuznetsova and Johnson (2017).

The stratified urn design (Wei, 1977, 1978a, 1978b) is the stratified biased coin randomisation with p depending on i. When $k = 2$ and $\pi_1 = \pi_2 = 1/2$, the fixed p in biased coin is replaced by a p_i depending on $D_{i-1}(\boldsymbol{z})$. According to Wei (1977), the urn design would force balance at the beginning of treatment allocation, and approach simple randomisation as the size of trial increases. A stratified urn design for general situation of $k \geq 3$ can be constructed using the method described in Zhao and Ramakrishnan (2016).

The previous three stratified covariate-adaptive randomisation schemes enforce balancedness of treatment assignment allocation across all strata, i.e., joint levels of $\boldsymbol{Z}$. However, the oldest covariate-adaptive randomisation scheme, the minimisation, is very different from these three methods.

First, consider $k = 2$ and $\pi_1 = \pi_2 = 1/2$. For each $i \in \{1, \ldots, n\}$, let $G_i^{(1)}$ be a weighted sum of squared or absolute differences between numbers of patients in two treatment arms over marginal levels of $\boldsymbol{Z}$, where the calculation is based on $i-1$ previously assigned patients and the assumption that the ith patient i is assigned to treatment 1, and let $G_i^{(2)}$ be the same sum except that the ith patient is assumed to be in treatment 2. For $a = 1$ or 2, $G_i^{(a)}$ represents the 'total amount of imbalance' in treatment numbers across the marginal levels of $\boldsymbol{Z}$ which exists if treatment a is assigned to the ith patient. Therefore, we would like to assign the ith patient by minimising $G_i^{(a)}$ over $a = 1, 2$, i.e., we assign the ith patient to treatment 1 if $G_i^{(1)} < G_i^{(2)}$, to 2 if $G_i^{(1)} > G_i^{(2)}$, and to 1 or 2 randomly if $G_i^{(1)} = G_i^{(2)}$. This is why the method is called the minimisation by Taves (1974). Pocock and Simon (1975) extended the minimisation by allowing minimisation with a given probability, i.e.,

$$P(A_i = e_1) = \begin{cases} p, & G_i^{(1)} < G_i^{(2)}, \\ 1/2, & G_i^{(1)} = G_i^{(2)}, \\ 1-p, & G_i^{(1)} > G_i^{(2)}, \end{cases}$$

where $p > 1/2$ is a fixed constant. Pocock and Simon's method is still referred to as the minimisation and Taves' minimisation is the special case with $p = 1$. For a general k and/or allocation, the minimisation can be similarly constructed (Han et al., 2009; Pocock & Simon, 1975).

If $\boldsymbol{Z}$ is one dimensional, then the minimisation is the same as the stratified biased coin method. For a multivariate $\boldsymbol{Z}$, the key distinction between the minimisation and the three previously described stratified randomisation methods is that enforcing treatment balancedness is at all joint levels of $\boldsymbol{Z}$ for the latter but only at marginal levels of $\boldsymbol{Z}$ for the former. For this reason, the minimisation is also called the marginal method in Ma et al. (2015) and Ye and Shao (2020). Enforcing treatment balance in marginal levels of $\boldsymbol{Z}$ is sufficient in most applications.

Any of the previously introduced covariate-adaptive randomisation schemes satisfy

(D1) $\{A_i, i = 1, \ldots, n\}$ and $\{Y_i^{(1)}, \ldots, Y_i^{(k)}, \boldsymbol{X}_i, i = 1, \ldots, n\}$ are conditionally independent given $\{\boldsymbol{Z}_i, i = 1, \ldots, n\}$.

Actually, (D1) almost always holds for covariate-adaptive randomisation, because treatments, not their assignments, may affect the potential responses as we discussed earlier for simple randomisation, and given $\boldsymbol{Z}_i$'s, the rest of $\boldsymbol{X}_i$'s contain covariate information not used in randomisation.

Furthermore, all covariate-adaptive randomisation schemes considered so far satisfy the following condition (D2) (Baldi Antognini & Zagoraiou, 2015). In what follows, $\Rightarrow$ denotes convergence in distribution as the sample size $n \to \infty$, and $\Rightarrow 0$ is in fact convergence to 0 in probability.

(D2) For every $i = 1, \ldots, n$, $P(A_i = e_a | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) = \pi_a$ and, for every a and every level $\boldsymbol{z}$ of $\boldsymbol{Z}$, $\{n(\boldsymbol{z})\}^{-1} D^{(a)}(\boldsymbol{z}) \Rightarrow 0$, where $D^{(a)}(\boldsymbol{z}) = n_a(\boldsymbol{z}) - \pi_a n(\boldsymbol{z})$, $n(\boldsymbol{z})$ is the number of patients with $\boldsymbol{Z}_i = \boldsymbol{z}$, and $n_a(\boldsymbol{z})$ is the number of patients with $\boldsymbol{Z}_i = \boldsymbol{z}$ under treatment a.

Note that $D^{(a)}(\boldsymbol{z})$ is a measure of the assignment imbalance in stratum $\boldsymbol{z}$. According to the asymptotic property of $D^{(a)}(\boldsymbol{z})$ in (D2), covariate-adaptive randomisation schemes can be classified into one of the following three types.

Type 1. For every a and every $\boldsymbol{z}$ of $\boldsymbol{Z}$, $\{n(\boldsymbol{z})\}^{-1/2} D^{(a)}(\boldsymbol{z}) \Rightarrow 0$.

Type 2. For every a, $D^{(a)}(\boldsymbol{z})$'s with all different strata $\boldsymbol{z}$'s are mutually independent and, for every $\boldsymbol{z}$, $\{n(\boldsymbol{z})\}^{-1/2} D^{(a)}(\boldsymbol{z}) \Rightarrow N(0, v_a)$, the normal distribution with mean 0 and a known variance $v_a > 0$.

Type 3. Methods not in Type 1 or 2.

The three types are in the order of the degree in enforcing the balancedness within every $\boldsymbol{z}$ using the

assignment imbalance measure $D^{(a)}(\boldsymbol{z})$. Type 1 is the strongest, requiring $D^{(a)}(\boldsymbol{z})$ diverging slower than the square root of within stratum $\boldsymbol{z}$ sample size. Representatives of Type 1 covariate-adaptive randomisation methods are stratified permuted block and biased coin schemes. In fact, under stratified permuted block randomisation, $D^{(a)}(\boldsymbol{z})$ is bounded; for the stratified biased coin method, it follows from a result in Efron (1971) that $D^{(a)}(\boldsymbol{z})$ is bounded in probability for every $\boldsymbol{z}$.

Type 2 is weaker than Type 1, as $\{n(\boldsymbol{z})\}^{-1/2}D^{(a)}(\boldsymbol{z})$ converges in distribution to $N(0, v_a)$, not 0. The stratified urn design is Type 2 with $v_a = 1/12$ when $k = 2$, $\pi_1 = \pi_2 = 1/2$ (Wei, 1978a, 1978b). Simple randomisation treated as a special case of covariate-adaptive randomisation is also Type 2. Finally, the minimisation is Type 3, since it is neither Type 1 nor Type 2 (Ye & Shao, 2020). Specifically, under minimisation, $D^{(a)}(\boldsymbol{z})$ and $D^{(a)}(\boldsymbol{z}')$ with $\boldsymbol{z} \neq \boldsymbol{z}'$ are not independent, and their relationship is complicated, because assignments are made according to marginal levels of $\boldsymbol{Z}$.

4. Validity and conservativeness of tests

Testing a null hypothesis of no treatment effect on potential outcomes is the most utilised statistical inference procedure in clinical trials. For a given null hypothesis H_0 and a significance level $\alpha > 0$, a test statistic T is a function of observed $\{Y_i, \boldsymbol{X}_i, i = 1, \ldots, n\}$, which is constructed such that H_0 is rejected if and only if T is outside of the interval $[z_{\alpha/2}, z_{1-\alpha/2}]$, where z_r is the rth quantile of a known distribution, usually the standard normal distribution, in which case H_0 is rejected if and only if $|T| > z_{1-\alpha/2}$ as $z_{1-\alpha/2} = -z_{\alpha/2}$. Here, we consider two sided tests; the discussion for a one sided test is similar and omitted. T is said to be asymptotically valid (or valid for simplicity) if

$$\sup_{P \text{ under } H_0} \lim_{n\to\infty} P\left(T \notin [z_{\alpha/2}, z_{1-\alpha/2}]\right) = \alpha \qquad (1)$$

T is said to be asymptotically conservative (or conservative for simplicity) if

$$\sup_{P \text{ under } H_0} \lim_{n\to\infty} P\left(T \notin [z_{\alpha/2}, z_{1-\alpha/2}]\right) < \alpha. \qquad (2)$$

4.1. Validity of conventional tests

As we discussed in Section 1, prior to 2010, there was almost no theoretical work and practitioners applied conventional tests developed under simple randomisation, which caused concerns about whether Type I error could be inflated. That is, if a conventional test T is applied after covariate-adaptive randomisation, does (1) still hold?

Forsythe (1987) concluded that a conventional test T still controls Type I error when $\boldsymbol{Z}$ used in minimisation is also included in the construction of T. However, this conclusion was based on simulation results under certain models.

The first piece of theoretical work in this area obtained by Shao et al. (2010) is that, under covariate-adaptive randomisation, a conventional T is valid in the sense of (1) if both of the following hold:

(i) The covariate $\boldsymbol{Z}$ used in covariate-adaptive rando misation is a function of all covariates used to construct the test T.
(ii) T is valid in the sense of (1) under any fixed set of treatment allocation $A_1, \ldots, A_n$.

Note that (i) coincides with Forsythe's simulation discovery. But (ii) requires the validity of T under any deterministic allocation $A_1, \ldots, A_n$, which can be realistically achieved only when a correct statistical model is used in constructing T. However, correctly impose a model is difficult. Although mathematically, (i)–(ii) is only sufficient not necessary for the validity of a conventional test T under covariate-adaptive randomisation, we can easily find an example in which T is not valid under covariate-adaptive randomisation when either (i) or (ii) fails; e.g., Shao et al. (2010) and Shao and Yu (2013).

4.2. Conservativeness of conventional tests

Before we answer Question (A) in Section 1 regarding the development of a valid test according to (1) under covariate-adaptive randomisation, we would like to address Question (B) in Section 1, i.e., whether or not a conventional test T is conservative in the sense of (2). If the answer is yes, then at least the Type I error is not inflated by using conventional tests.

The first result of this kind was obtained by Shao et al. (2010) regarding the two sample t-test under a homogeneous one-way analysis of covariance model. The result is, the conventional two sample t-test for comparing two treatments ($k = 2$) is conservative according to (2) under the stratified biased coin randomisation. Following this work, results about the conservativeness of different conventional tests under different models and covariate-adaptive randomisation methods have been obtained by Hu and Hu (2012), Shao and Yu (2013), Ma et al. (2015), Bugni et al. (2018), Ye (2018), and Ye and Shao (2020). In particular, under Type 1 or 2 randomisation schemes described in Section 3, Ye and Shao (2020) proved the conservativeness of the conventional log-rank and score tests for survival analysis, which is a substantial advance in the theory of this area. Unfortunately, no result is available for the conservativeness of conventional tests under minimisation, except for some unrealistic cases. Furthermore, the available results are for particular conventional tests, i.e., no general result is available.

The reason why conventional tests become conservative under some covariate-adaptive randomisation schemes as well as why the result is not available for minimisation can be explained as follows. Many (if not most) conventional tests are ratios with numerators being statistics accessing the plausibility of the null hypothesis H_0 and denominators being standard errors estimating the standard deviations of the corresponding numerators. For example, the two sample t-test for testing effect between two treatments ($k = 2$) is

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \tag{3}$$

where n_a is the number of patients assigned to treatment a, $\overline{Y}_a$ and S_a^2 are the sample mean and sample variance, respectively, based on Y_i's under treatment a; the numerator $\overline{Y}_1 - \overline{Y}_2$ of T in (3) accesses the plausibility of the null hypothesis $H_0 : E(Y^{(1)} - Y^{(2)}) = 0$, and the denominator $\sqrt{S_1^2/n_1 + S_2^2/n_2}$ estimates the asymptotic standard deviation of $\overline{Y}_1 - \overline{Y}_2$. Under a Type 1 or 2 covariate-adaptive randomisation scheme, it is usually true that the numerator of T in (3) still measures the plausibility of H_0, and the denominator of T in (3) is too large because the Type 1 or 2 covariate-adaptive randomisation scheme typically reduces the variation of numerator after enforcing the balancedness of treatment assignments. Specifically, $\overline{Y}_1$ and $\overline{Y}_2$ are independent under simple randomisation but are negatively correlated under Type 1 or Type 2 covariate-adaptive randomisation and, consequently, the variance of $\overline{Y}_1 - \overline{Y}_2$ is smaller under covariate-adaptive randomisation and $S_1^2/n_1 + S_2^2/n_2$ still estimates the variance of $\overline{Y}_1 - \overline{Y}_2$ under simple randomisation. The reduction in variation together with the fact that the denominator of conventional test does not account for this reduction lead to the conservativeness of conventional test.

As we discussed in Section 3, the stratified permuted block and biased coin randomisation schemes are Type 1 and the stratified urn designs are Type 2. Hence conventional tests are conservative under these randomisation schemes.

The minimisation, however, is neither Type 1 nor Type 2 (Ye & Shao, 2020). The only available result on the asymptotic distribution of $\overline{Y}_1 - \overline{Y}_2$ under minimisation is obtained (Ma et al., 2015) under a very restrictive and nearly unrealistic condition that not only the relationship between the observed response Y_i and $\boldsymbol{Z}_i$ is linear, but also all components of $\boldsymbol{Z}_i$ are independent and there is no other covariate in the linear model. Because the minimisation only enforces the marginal balancedness of treatment assignments, its asymptotic properties are very complicated and a general result about the asymptotic distribution of a simple statistic like $\overline{Y}_1 - \overline{Y}_2$ is not available. Some progress has been made in some recent work (Hu & Zhang, 2020), but the problem is not completely solved.

4.3. Development of valid tests

We now return to address Question (A) in Section 1. Although a conservative test controls the Type I error rate, it may lose power of the test and, thus, may not be appreciated by clinical trialists.

From the discussion in Section 4.1, a conventional test is valid according to Equation (1) if (i)–(ii) hold, but (ii) requires prefect modelling that may be unrealistic, since model misspecification often occurs especially when there are many covariates. The discussion in Section 4.2 actually suggests that we modify the denominator of a conventional test to develop a valid test under covariate-adaptive randomisation. The first result was also obtained by Shao et al. (2010) who proposed a bootstrap variance estimator for the two sample t-test with a component of re-generating treatment assignments in every bootstrap sample to account for the correct variation under the stratified biased coin randomisation. The resulting bootstrap test replaces the denominator of two sample t-test in (3) by the squared root of the bootstrap variance estimator and is valid according to (1). This bootstrap method can be extended to modifying many other conventional tests, for Type 1 or 2 covariate-adaptive randomisation scheme (Shao & Yu, 2013; Ye & Shao, 2020).

With some effort on deriving the asymptotic distribution of the numerator of a conventional test under Type 1 or 2 covariate-adaptive randomisation, a valid test can also be constructed by correctly estimating the asymptotic variance of the numerator (Ye, 2018; Ye & Shao, 2020). For the conventional two sample t-test in (3), for example, Ye (2018) showed that a valid test under stratified biased coin randomisation can be obtained by replacing the denominator of the t-test by $2\sqrt{\sum_{\boldsymbol{z}} n(\boldsymbol{z})S^2(\boldsymbol{z})}/n$, where $S^2(\boldsymbol{z})$ is the sample variance based on Y_i's in stratum $\boldsymbol{Z} = \boldsymbol{z}$. Compared with the bootstrap, this approach does not require a large amount of computation and has another advantage to be discussed later.

Perhaps a better approach is to directly derive a valid test based on a given covariate-adaptive randomisation scheme or a general group of randomisation schemes. This will be discussed in Section 5 when we consider general inference procedures.

Another stream of methods is based on re-randomisation or permutation, e.g., Simon and Simon (2011), Kaiser (2012), and Bugni et al. (2018). In the rest of this section, we discuss in details about the re-randomisation approach in Simon and Simon (2011), which is somewhat similar to the bootstrap method. Consider $k = 2$ and $H_0 : Y^{(1)} \sim Y^{(2)}$. Under H_0, $Y^{(1)}$ and $Y^{(2)}$ are exchangeable so that we create potential outcome $\widetilde{Y}_i^{(1)} = \widetilde{Y}_i^{(2)} = Y_i$ for patient

i. Let $\mathcal{A} = (A_1, \ldots, A_n)$ be the observed treatment assignments under the given covariate-adaptive randomisation scheme. Any test T can be written as $T(\mathcal{A}, \mathcal{O})$, where $\mathcal{O} = \{\widetilde{Y}_i^{(1)}, \widetilde{Y}_i^{(2)}, \boldsymbol{X}_i, i = 1, \ldots, n\}$. Let $\mathcal{C} = (C_1, \ldots, C_n)$ be randomly generated treatment assignments under the same randomisation scheme, i.e., $\mathcal{C} \sim \mathcal{A}$ conditioned on $\boldsymbol{Z}$, $T(\mathcal{C}, \mathcal{O})$ be $T(\mathcal{A}, \mathcal{O})$ with $\mathcal{A}$ replaced by $\mathcal{C}$, and let $F_{\mathcal{O}}$ be the cumulative conditional distribution function of $T(\mathcal{C}, \mathcal{O})$ given $\mathcal{O}$. From the probability theory,

$$P\Big\{T(\mathcal{C}, \mathcal{O}) < F_{\mathcal{O}}^{-1}(\alpha/2) \text{ or } T(\mathcal{C}, \mathcal{O}) > F_{\mathcal{O}}^{-1}(1-\alpha/2)\Big|\mathcal{O}\Big\} \le \alpha.$$

Hence, unconditionally, under H_0,

$$P\Big\{T(\mathcal{C}, \mathcal{O}) < F_{\mathcal{O}}^{-1}(\alpha/2) \text{ or } T(\mathcal{C}, \mathcal{O}) > F_{\mathcal{O}}^{-1}(1-\alpha/2)\Big\} \le \alpha$$

and if we reject H_0 if and only if T is outside of the interval $[F_{\mathcal{O}}^{-1}(\alpha/2), F_{\mathcal{O}}^{-1}(1-\alpha/2)]$, then this T has Type I error rate $\le \alpha$ for every n.

Two issues remain to be considered. The first one is that the quantile $F_{\mathcal{O}}^{-1}(r)$ usually has no explicit form and approximation such as Monte Carlo is needed. The second issue is that this method may be conservative for every n, because $T(\mathcal{C}, \mathcal{O})$ with random $\mathcal{C}$ is discrete. At this stage, it is still unknown whether result (1) holds for this method, since the previous argument shows that the left-hand side of (1) $\le \alpha$, but we cannot prove the equality in (1) holds, i.e., we cannot rule out the possibility that (2) actually holds so that the re-randomisation method is conservative.

4.4. Tests in survival analysis

We review some available theory for survival analysis, since covariate-adaptive randomisation has a long history of application in survival trials. In fact, all 7 articles in the 2018 *New England Journal of Medicine* cited in Section 1 are about survival trials.

For simplicity, we focus on the case of $k = 2$.

The data structure for survival analysis is described in the beginning of Section 2, where the potential outcome $Y^{(a)} = \min(T^{(a)}, C^{(a)})$, $T^{(a)}$ is the potential survival, and $C^{(a)}$ is the potential censoring, under treatment a. It is typically assumed that conditional on covariate $\boldsymbol{X}$, $T^{(a)}$ and $C^{(a)}$ are independent and the ratio $P(C^{(1)} \ge t|\boldsymbol{X})/P(C^{(2)} \ge t|\boldsymbol{X})$ is a function of t only.

The most common analysis in survival trials is testing whether two treatments have different effect on the conditional distributions of $T^{(a)}$ given $\boldsymbol{X}$. Let $\lambda(t, \boldsymbol{x}, a)$ be the underlying hazard function of $T^{(a)}$ given $\boldsymbol{X} = \boldsymbol{x}$, $a = 1, 2$. The null hypothesis of interest is H_0 : $\lambda(t, \boldsymbol{x}, 1) = \lambda(t, \boldsymbol{x}, 2)$ for all possible t and $\boldsymbol{x}$.

Without imposing any model, a conventional nonparametric test for H_0 is the log-rank test

$$T = \sum_{i=1}^{n} \int_0^\infty \left\{A_i - \frac{S_1(t)}{S(t)}\right\} \mathrm{d}N_i(t) \times \left[\sum_{i=1}^{n} \int_0^\infty \frac{S_1(t)S_2(t)}{\{S(t)\}^2} \mathrm{d}N_i(t)\right]^{-1/2}, \quad (4)$$

where $S_a(t) = \sum_{i=1}^n I(A_i = e_a) I(Y_i^{(a)} \ge t)$, $I(C)$ is the indicator of event C, $S(t) = S_1(t) + S_2(t)$, $N_i(t) = I(A_i = e_1)N_i^{(1)}(t) + I(A_i = e_2)N_i^{(2)}(t)$, and $N_i^{(a)}(t) = I(T_i^{(a)} \le C_i^{(a)}) I(Y_i^{(a)} \le t)$, $a = 1, 2$. Similar to the two sample t-test in (3), the log-rank test that is valid according to (1) under simple randomisation is conservative in the sense of (2) under Type 1 or 2 covariate-adaptive randomisation, because the denominator of T in (4) is too large as a standard error for the numerator of T. A valid modified log-rank test is derived by replacing the denominator of T with the squared root of a stratified variance estimator given in Formula (20) of Ye and Shao (2020).

In survival analysis, the following Cox proportional hazard model is very popular:

$$\lambda(t, \boldsymbol{x}, a) = \lambda_0(t) \exp(\theta a + \beta^{\mathrm{T}} \boldsymbol{x}), \quad (5)$$

where θ is an unknown parameter, β^{T} is the transpose of a vector β of unknown parameters, and $\lambda_0(t)$ is an unspecified baseline hazard function. If the Cox model is correct, then the null hypothesis is the same as $H_0 : \theta = 0$, and a score test of H_0 can be derived using the partial likelihood under the Cox model. The idea is that the score test is more powerful than the log-rank test if the Cox model is correct. Even if the Cox model could be misspecified, it can be used as a working model under the model-assisted approach, i.e., a model is used to assist the derivation of an inference procedure that is efficient when the model is correct and is still asymptotically valid when the model is incorrect.

Under simple randomisation, a valid model-assisted score test was derived (DiRienzo & Lagakos, 2002; Kong & Slud, 1997; Lin & Wei, 1989), which is often more powerful than the log-rank test in (4) without using any covariates. This conventional score test, however, is shown in Ye and Shao (2020) to be conservative under Type 1 or 2 covariate-adaptive randomisation, because of the same reason that the denominator of the score test is too large as a standard error. Again, we can obtain a valid score test by replacing the denominator with the squared root of a stratified variance estimator (Ye & Shao, 2020).

We can also apply the bootstrap or re-randomisation discussed in Section 4.3 to construct valid tests. However, the bootstrap or re-randomisation discussed in Section 4.3 is not correct in survival analysis, unless we assume $P(C^{(1)} \ge t|\boldsymbol{X}) = P(C^{(0)} \ge t|\boldsymbol{X})$ for all t.

The reason is that, to apply the bootstrap or re-randomisation, the observed $(Y_i, \boldsymbol{X}_i)$'s have to be exchangeable across i under H_0. Under H_0, although $T_i^{(a)}$'s are exchangeable, $C_i^{(a)}$'s are not unless $P(C^{(1)} \geq t|\boldsymbol{X}) = P(C^{(2)} \geq t|\boldsymbol{X})$ for all t. Even if the treatment has no effect on the potential survival time, it may have some effect on the potential censoring due to some practical reasons.

5. Valid inference

We have already discussed to some extent how to construct valid tests under covariate-adaptive randomisation. There are a few shortcomings in those available results reviewed in Section 4. First, an obvious one is that some results/methods rely on correct specification of a model. Second, all results/methods in Section 4 depend on covariate-adaptive randomisation schemes; in particular, Type 1 or 2 randomisation method is required, which excludes the minimisation. Third, only testing hypotheses is considered, not other inference such as confidence sets. Finally, all methods in Section 4 are modifications of conventional procedures.

In this section, we would like to address the following re-phrased Question (A) in Section 1:

(A) Can we develop an inference procedure valid under covariate-adaptive randomisation with very little model assumption?

5.1. Testing in survival analysis

We begin with the log-rank test for survival data in the case of $k = 2$. The stratified log-rank test (Peto et al., 1976) is simply the log-rank test in (4) stratified with all levels of the discrete covariate $\boldsymbol{Z}$ utilised in covariate-adaptive randomisation:

$$T = \sum_{\boldsymbol{z}} \sum_{i \in L(\boldsymbol{z})} \int_0^\infty \left\{ A_i - \frac{S_1(t,\boldsymbol{z})}{S(t,\boldsymbol{z})} \right\} \mathrm{d}N_i(t) \times \left[\sum_{\boldsymbol{z}} \sum_{i \in L(\boldsymbol{z})} \int_0^\infty \frac{S_1(t,\boldsymbol{z}) S_2(t,\boldsymbol{z})}{\{S(t,\boldsymbol{z})\}^2} \mathrm{d}N_i(t) \right]^{-1/2}, \tag{6}$$

where $L(\boldsymbol{z})$ is the stratum of patients with $\boldsymbol{Z}_i = \boldsymbol{z}$, $S_a(t,\boldsymbol{z}) = \sum_{i \in L(\boldsymbol{z})} I(A_i = e_a) I(Y_i^{(a)} \geq t)$, and $S(t,\boldsymbol{z}) = S_1(t,\boldsymbol{z}) + S_2(t,\boldsymbol{z})$. Although the stratified log-rank test in (6) exhibits nice empirical properties under covariate-adaptive randomisation (Lachin et al., 1988; Xu et al., 2016) and has been used for a long time, the first proof of its validity according to (1) comes from Ye and Shao (2020) with some efforts. The proof actually shows that the stratified log-rank test is valid for any covariate-adaptive randomisation method, including the minimisation, as long as the minimal conditions (D1) –(D2) are satisfied.

Why does stratification make so much difference? Recall that in Section 4.1 we comment that a test will be valid if two conditions are satisfied: (i) $\boldsymbol{Z}$ used in randomisation is also used in constructing the test and (ii) a correct model is used to derive the test. Note that the stratification with strata being levels of $\boldsymbol{Z}$ can be viewed as a kind of modelling based on the discrete covariate $\boldsymbol{Z}$, and such modelling is always correct. Thus (ii) has been met if we stratify using $\boldsymbol{Z}$. To meet (i), we must fully stratify, i.e., use all strata defined by joint levels of $\boldsymbol{Z}$, not partially stratify. It can be shown that if we combine some strata in the construction of the stratified log-rank test, then the resulting test is not valid.

The only issue with the stratified log-rank test in (6) is that it is not efficient if $\boldsymbol{Z} \neq \boldsymbol{X}$, i.e., $\boldsymbol{X}$ contains more information than $\boldsymbol{Z}$. In fact, we cannot definitely tell whether the stratified log-rank test is more powerful than the unstratified log-rank test in (4) under simple randomisation, which is similar to the issue of a stratified sample mean may not be always more efficient than the unstratified sample mean in survey sampling. Ye and Shao (2020) showed by simulation that a modified log-rank test that replaces the denominator of T in (4) with a stratified standard error may be more powerful than the stratified log-rank test in (6). The efficiency issue will be further considered in Section 6.

5.2. Inference on average or quantile treatment effect

Next, we consider inference on the population mean difference $\theta = E(Y^{(a)} - Y^{(b)})$ with any two fixed treatments a and b in a trial with $k \geq 2$ treatment arms. As the development of inference procedures often starts with finding estimators of the parameter of interest, we first review some available estimators of θ.

The simplest estimator of θ is the sample mean difference $\overline{Y}_a - \overline{Y}_b$, where $\overline{Y}_a$ is the sample mean of Y_i's under treatment $a = 1, \ldots, k$. Bugni et al. (2018) proposed another estimator called the strata fixed effect estimator in their Section 4.2. The asymptotic distributions of $\overline{Y}_a - \overline{Y}_b$ and the strata fixed effect estimator have been derived under Type 1 or 2 covariate-adaptive randomisation, but they are not available for Type 3 covariate-adaptive randomisation such as minimisation due to the lack of theory on Type 3 methods.

The following post-stratified estimator of θ, similar to the stratified log-rank test in (6), is proposed by Bugni et al. (2019) and Ye et al. (2020),

$$\widehat{\theta}_S = \sum_{\boldsymbol{z}} \frac{n(\boldsymbol{z})}{n} \{\overline{Y}_a(\boldsymbol{z}) - \overline{Y}_b(\boldsymbol{z})\}, \tag{7}$$

where $\overline{Y}_a(\boldsymbol{z})$ is the sample mean of Y_i's from patients in post-stratum $L(\boldsymbol{z})$ under treatment $a = 1, \ldots, k$. If the weight $n(z)/n$ in (7) is replaced by the population weight $P(\boldsymbol{Z} = \boldsymbol{z})$, then $\widehat{\theta}_S$ in (7) is the stratified estimator in survey sampling. Since $P(\boldsymbol{Z} = \boldsymbol{z})$ is substituted

by $n(z)/n$ and $L(\boldsymbol{z})$ is formed after $\boldsymbol{Z}$ is observed, the estimator $\widehat{\theta}_S$ is referred to as post-stratified estimator in survey sampling.

Applying different techniques, Bugni et al. (2019) and Ye et al. (2020) independently established that, if (C1) and (D1)–(D2) hold and the second order moments of $Y^{(a)}$ and $Y^{(b)}$ are finite, then

$$\sqrt{n}(\widehat{\theta}_S - \theta) \Rightarrow N(0, \sigma_S^2), \tag{8}$$

where

$$\sigma_S^2 = E\left\{\text{var}(Y^{(a)}|\boldsymbol{Z})/\pi_a + \text{var}(Y^{(b)}|\boldsymbol{Z})/\pi_b\right\} + \text{var}\{E(Y^{(a)} - Y^{(b)}|\boldsymbol{Z})\}.$$

Result (8) is model free, i.e., only (C1) and the second-order moments of the potential outcomes are required. It is applicable to any covariate-adaptive randomisation method satisfying (D1)–(D2), most noticeably the minimisation for which very little is known about its theoretical property, as the minimisation is neither Type 1 nor Type 2. Another interesting fact is that the limiting variance σ_S^2 is invariant with respect to randomisation methods. Hence, not only result (8) holds for any covariate-adaptive randomisation method as long as the minimal (D1) –(D2) are satisfied, but also $\widehat{\theta}_S$ in (7) has the same asymptotic distribution and efficiency regardless of which randomisation scheme is used for treatment assignments. Such kind of result has not be seen in the literature except that Ye and Shao (2020) showed that the asymptotic distribution of the stratified log-rank test in (6) is invariant to the randomisation schemes. Existing results in the literature (Bugni et al., 2018; Ma et al., 2015; Shao & Yu, 2013; Shao et al., 2010) are typically dependent with randomisation methods and many of them are not applicable to Type 3 methods such as the minimisation.

When the covariate-adaptive randomisation scheme is Type 1, result (8) also holds with $\widehat{\theta}_S$ replaced by the strata fixed effect estimator in Bugni et al. (2018). In general, however, $\widehat{\theta}_S$ is asymptotically more efficient than the strata fixed effect estimator or the simple estimator $\overline{Y}_a - \overline{Y}_b$.

For inference on θ under any type covariate-adaptive randomisation, if $\widehat{\theta}_S$ is adopted to estimate θ, then all we need to do is to derive an estimator $\widehat{\sigma}_S^2$ of σ_S^2 that is consistent, i.e., $\widehat{\sigma}_S^2 - \sigma_S^2 \Rightarrow 0$ under any type covariate-adaptive randomisation. This is actually not difficult once we establish a result like (8). It is shown in Ye et al. (2020) that a consistent estimator of σ_S^2 under any type covariate-adaptive randomisation is

$$\widehat{\sigma}_S^2 = \frac{1}{n}\sum_{\boldsymbol{z}} n^2(\boldsymbol{z})\left\{\frac{S_a^2(\boldsymbol{z})}{n_a(\boldsymbol{z})} + \frac{S_b^2(\boldsymbol{z})}{n_b(\boldsymbol{z})}\right\} + \frac{1}{n}\sum_{\boldsymbol{z}} n(\boldsymbol{z})\left\{\overline{Y}_a(\boldsymbol{z}) - \overline{Y}_b(\boldsymbol{z})\right\}^2 - \widehat{\theta}_S^2,$$

where $n_a(\boldsymbol{z})$ and $S_a^2(\boldsymbol{z})$ are the sample size and sample variance of Y_i's, respectively, of the patients in stratum $\boldsymbol{Z} = \boldsymbol{z}$ and under treatment a.

Under any randomisation scheme satisfying (D1)–(D2), an asymptotically valid $(1-\alpha)\%$ confidence interval for θ has limits $\widehat{\theta}_S \pm z_{1-\alpha/2}\widehat{\sigma}_S$, where $z_{1-\alpha}$ is the quantile of the standard normal distribution.

More estimators of the average treatment effect θ are considered in Section 6.

We now consider inference on another important parameter, the quantile treatment effect defined as $q_\tau^{(a)} - q_\tau^{(b)}$ in Section 2, where $q_\tau^{(a)}$ is the τth quantile of the distribution of $Y^{(a)}$ under treatment a and τ is a fixed fraction.

Unlike the means, for quantiles we cannot use differences as in (7). Instead, we estimate $q_\tau^{(a)}$ and $q_\tau^{(b)}$ separately, and then take a difference of estimates. Under treatment a, we estimate the marginal distribution of $Y^{(a)}$ at a fixed point y as

$$\widehat{F}^{(a)}(y) = \frac{1}{n}\sum_{\boldsymbol{z}} \frac{n(\boldsymbol{z})}{n_a(\boldsymbol{z})} \sum_{i\in L(\boldsymbol{z})} I(A_i = e_a) I(Y_i^{(a)} \le y), \tag{9}$$

$a = 1, \ldots, k$. Then, $q_\tau^{(a)}$ is estimated by $\widehat{q}_\tau^{(a)} =$ the τth quantile of $\widehat{F}^{(a)}$, and $q_\tau^{(a)} - q_\tau^{(b)}$ is estimated as $\widehat{q}_\tau^{(a)} - \widehat{q}_\tau^{(b)}$. For inference on quantiles, however, a simple estimator of the asymptotic variance of $\widehat{q}_\tau^{(a)}$ may not be easily obtained. Methods such as the bootstrap or Woodruff's interval may be applied (Shao, 2003).

The stratification in (6), (7) or (9), together with the asymptotic theory, provides a solid foundation for valid and model free inference after covariate-adaptive randomisation and, thus, it largely eliminates the concern and controversy as discussed by regulatory agencies about the use of covariate-adaptive randomisation such as minimisation.

Combining the results and discussions in this section and Section 5.1, we reach a general conclusion that a valid inference procedure can be obtained as long as the covariate $\boldsymbol{Z}$ utilised in covariate-adaptive randomisation is fully used in the construction of inference procedure. A simple way to do this is to use all joint levels of $\boldsymbol{Z}$ as strata.

It can be seen that the conditions needed for this conclusion is much weaker than (i) and (ii) stated in Section 4.1, but (i)–(ii) in Section 4.1 are considered for the validity of a conventional test under covariate-adaptive randomisation.

5.3. Effect of types of randomisation schemes

Result (8) about the asymptotic distribution of $\widehat{\theta}_S$ in (7) is invariant to any types of randomisation schemes described in Section 3. But this does not imply that the stratification in (7) or in (6) is the best way for inference, especially when problems other than the inference on

average treatment effect are considered. An example is that the modified log-rank test in Ye and Shao (2020) may be more powerful than the stratified log-rank test in (6), as discussed in the end of Section 5.1.

If an inference procedure is not invariant to different randomisation schemes, then it is interesting to find out which randomisation scheme, or which type, provides better inference procedures. For the modified log-rank test in Ye and Shao (2020), it is more powerful when a Type 1 randomisation scheme is used, rather than the Type 2 or 3. The same may be true for any inference procedure not invariant to different randomisation schemes. For different Type 1 methods, such as the stratified permuted block and the stratified biased coin methods, so far there is no result indicating that the inference procedures based on these two randomisation schemes have different performances.

6. Efficiency considerations

Question (C) in Section 1 is about whether a test under covariate-adaptive randomisation can be more powerful than it is under simple randomisation. Another question is, if $\boldsymbol{Z}$ is used in randomisation and stratification as in (6) or (7) and if $\boldsymbol{X}$ contains more information than $\boldsymbol{Z}$, can we obtain more powerful tests or more efficient estimators by utilising covariate information in $\boldsymbol{X}$ that is not in $\boldsymbol{Z}$? Note that $\boldsymbol{X}$ may contain a component that is not in $\boldsymbol{Z}$ but is related with the potential responses $Y^{(1)}, \ldots, Y^{(k)}$, or some components of $\boldsymbol{Z}$ are discretised components of $\boldsymbol{X}$ and the remaining information after discretisation is still useful in predicting the potential responses.

6.1. Adjusting for covariates

We first consider the second question in the estimation of $\theta = E(Y^{(a)} - Y^{(b)})$ for two fixed treatments a and b. Let $\boldsymbol{U}$ be a function of $\boldsymbol{X}$ that we want to further utilise in improving the efficiency of $\widehat{\theta}_S$ in (7). Since the information generated by $\boldsymbol{Z}$ is not in that of $\boldsymbol{U}$, we assume that $\mathrm{var}(\boldsymbol{U}|\boldsymbol{Z} = \boldsymbol{z})$ is positive definite for every $\boldsymbol{z}$.

For model free estimation and inference, we do not want to impose any model between the potential responses and $\boldsymbol{U}$. In fact, it is hard to find a correct model within each stratum $\boldsymbol{Z} = \boldsymbol{z}$, if we still apply stratification in estimating θ. How do we adjust for covariates without using a model? Ye et al. (2020) adopted the model-assisted generalised regression approach in survey sampling, first discussed in Cassel et al. (1976) and studied extensively in the literature, for example, Särndal et al. (2003), Shao and Wang (2014), and Ta et al. (2020).

In this section, we review some results from Ye et al. (2020). Let $\boldsymbol{U}_i$ be the covariate $\boldsymbol{U}$-value of patient i, and for each $\boldsymbol{z}$, let $\overline{\boldsymbol{U}}_a(\boldsymbol{z})$ be the sample mean of $\boldsymbol{U}_i$'s of patients in stratum $L_a(\boldsymbol{z}) = \{i : \boldsymbol{Z}_i = \boldsymbol{z}$ under treatment $a\}$, and

$$\widehat{\beta}_a(\boldsymbol{z}) = \left[\sum_{i\in L_a(\boldsymbol{z})} \{\boldsymbol{U}_i - \overline{\boldsymbol{U}}_a(\boldsymbol{z})\}\{\boldsymbol{U}_i - \overline{\boldsymbol{U}}_a(\boldsymbol{z})\}^{\mathrm{T}}\right]^{-1} \times \sum_{i\in L_a(\boldsymbol{z})} \{\boldsymbol{U}_i - \overline{\boldsymbol{U}}_a(\boldsymbol{z})\}Y_i.$$

Within treatment a and stratum $L(\boldsymbol{z}) = \{\boldsymbol{Z} = \boldsymbol{z}\}$, $\widehat{\beta}_a(\boldsymbol{z})$ is the least squares estimator of the coefficient vector in front of $\boldsymbol{U}$ under a linear model between $Y^{(a)}$ and $\boldsymbol{U}$, but the model is not required to be correct. An estimator of θ following $\widehat{\theta}_S$ but further adjusting for covariate $\boldsymbol{U}$ is (Ye et al., 2020)

$$\widehat{\theta}_A = \sum_{\boldsymbol{z}} \frac{n(\boldsymbol{z})}{n}[\overline{Y}_a(\boldsymbol{z}) - \overline{Y}_b(\boldsymbol{z}) - \{\overline{\boldsymbol{U}}_a(\boldsymbol{z}) - \overline{\boldsymbol{U}}(\boldsymbol{z})\}^{\mathrm{T}}\widehat{\beta}_a(\boldsymbol{z}) + \{\overline{\boldsymbol{U}}_b(\boldsymbol{z}) - \overline{\boldsymbol{U}}(\boldsymbol{z})\}^{\mathrm{T}}\widehat{\beta}_b(\boldsymbol{z})],$$

where $\overline{\boldsymbol{U}}(\boldsymbol{z})$ is the sample mean of $\boldsymbol{U}_i$'s of all patients in stratum $L(\boldsymbol{z})$.

An alternative estimator $\widehat{\theta}_B$ of θ in Ye et al. (2020) is obtained by replacing both $\widehat{\beta}_a(\boldsymbol{z})$ and $\widehat{\beta}_b(\boldsymbol{z})$ in the definition of $\widehat{\theta}_A$ with a combined estimator

$$\widehat{\beta}(\boldsymbol{z}) = \left[\sum_{a=1}^{k}\sum_{i\in L(\boldsymbol{z}),A_i=a} \{\boldsymbol{U}_i - \overline{\boldsymbol{U}}_a(\boldsymbol{z})\}\{\boldsymbol{U}_i - \overline{\boldsymbol{U}}_a(\boldsymbol{z})\}^{\mathrm{T}}\right]^{-1} \times \sum_{a=1}^{k}\sum_{i\in L(\boldsymbol{z}),A_i=a} \{\boldsymbol{U}_i - \overline{\boldsymbol{U}}_a(\boldsymbol{z})\}Y_i.$$

When $k > 2$, both $\overline{\boldsymbol{U}}(z)$ and $\widehat{\beta}(\boldsymbol{z})$ involve data from patients in treatment arms other than a and b.

The following result parallel to result (8) is established in Ye et al. (2020). If (C1) and (D1)–(D2) hold and the second order moments of $Y^{(a)}$ and $\boldsymbol{U}$ are finite, then

$$\sqrt{n}(\widehat{\theta}_A - \theta) \Rightarrow N(0, \sigma_A^2) \quad \text{and} \quad \sqrt{n}(\widehat{\theta}_B - \theta) \Rightarrow N(0, \sigma_B^2), \tag{10}$$

where

$$\begin{aligned}\sigma_A^2 = {} & E[\mathrm{var}\{Y^{(a)} - \boldsymbol{U}^{\mathrm{T}}\beta_a(\boldsymbol{Z})|\boldsymbol{Z}\}/\pi_a \\ & + \mathrm{var}\{Y^{(b)} - \boldsymbol{U}^{\mathrm{T}}\beta_b(\boldsymbol{Z})|\boldsymbol{Z}\}/\pi_b] \\ & + E[\{\beta_a(\boldsymbol{Z}) - \beta_b(\boldsymbol{Z})\}^{\mathrm{T}} \\ & \quad \times \mathrm{var}(\boldsymbol{U}|\boldsymbol{Z})\{\beta_a(\boldsymbol{Z}) - \beta_b(\boldsymbol{Z})\}] \\ & + \mathrm{var}\{E(Y^{(a)} - Y^{(b)}|\boldsymbol{Z})\}, \\ \sigma_B^2 = {} & E[\mathrm{var}\{Y^{(a)} - \boldsymbol{U}^{\mathrm{T}}\beta(\boldsymbol{Z})|\boldsymbol{Z}\}/\pi_a \\ & + \mathrm{var}\{Y^{(b)} - \boldsymbol{U}^{\mathrm{T}}\beta(\boldsymbol{Z})|\boldsymbol{Z}\}/\pi_b] \\ & + \mathrm{var}\{E(Y^{(a)} - Y^{(b)}|\boldsymbol{Z})\},\end{aligned}$$

$\beta_a(\boldsymbol{z}) = \{\text{var}(\boldsymbol{U}|\boldsymbol{Z}=\boldsymbol{z})\}^{-1}\text{cov}(\boldsymbol{U}, Y^{(a)}|\boldsymbol{Z}=\boldsymbol{z})$, $a = 1, \ldots, k$, and $\beta(\boldsymbol{z}) = \sum_{a=1}^{k} \pi_a\beta_a(\boldsymbol{z})$.

Several conclusions can be made from result (10). First, result (10) is model free and invariant with respect to covariate-adaptive randomisation schemes, as long as the minimal (D1)–(D2) hold.

Second, from the definitions of σ_S^2 and σ_A^2, it is shown in Ye et al. (2020) that

$$\begin{aligned}
&\sigma_S^2 - \sigma_A^2 \\
&= E\Big[\{\pi_b\beta_a(\boldsymbol{Z}) + \pi_a\beta_b(\boldsymbol{Z})\}^{\mathrm{T}} \\
&\qquad \times \text{var}(\boldsymbol{U} \mid \boldsymbol{Z})\{\pi_b\beta_a(\boldsymbol{Z}) + \pi_a\beta_b(\boldsymbol{Z})\}\Big] \\
&\qquad \times \{\pi_a\pi_b(\pi_a + \pi_b)\}^{-1} \\
&+ E\Big[\{\beta_a(\boldsymbol{Z}) - \beta_b(\boldsymbol{Z})\}^{\mathrm{T}}\,\text{var}(\boldsymbol{U} \mid \boldsymbol{Z})\,\{\beta_a(\boldsymbol{Z}) - \beta_b(\boldsymbol{Z})\}\Big] \\
&\qquad \times \{(\pi_a + \pi_b)^{-1} - 1\}
\end{aligned}$$

and, hence, adjusting covariate $\boldsymbol{U}$ always gains efficiency, i.e., $\widehat{\theta}_A$ is asymptotically more efficient than $\widehat{\theta}_S$, unless

$$\begin{aligned}
&\pi_b\beta_a(\boldsymbol{z}) + \pi_a\beta_b(\boldsymbol{z}) = 0 \quad \text{and} \\
&\{\beta_a(\boldsymbol{z}) - \beta_b(\boldsymbol{z})\}(1 - \pi_a - \pi_b) = 0 \quad \text{for every } \boldsymbol{z},
\end{aligned} \tag{11}$$

in which case $\widehat{\theta}_S$ and $\widehat{\theta}_A$ have the same asymptotic efficiency. When there are more than two treatments, $1 - \pi_a - \pi_b > 0$ and, consequently, (11) holds only when $\beta_a(\boldsymbol{z}) = \beta_b(\boldsymbol{z}) = 0$ for every $\boldsymbol{z}$, i.e., $\boldsymbol{U}$ is uncorrelated with the potential responses $Y^{(a)}$ and $Y^{(b)}$ after conditioning on $\boldsymbol{Z}$ so that adjusting for $\boldsymbol{U}$ is unnecessary. When there are only two treatments, (11) also holds if $\pi_a = \pi_b = 1/2$ and $\beta_a(\boldsymbol{z}) = -\beta_b(\boldsymbol{z})$ for every $\boldsymbol{z}$.

Third, from the definitions of σ_A^2 and σ_B^2, it can be shown (Ye et al., 2020) that

$$\begin{aligned}
&\sigma_B^2 - \sigma_A^2 \\
&= E\Big[\{\beta_a(\boldsymbol{Z}) - \beta(\boldsymbol{Z})\}^{\mathrm{T}}\text{var}(\boldsymbol{U} \mid \boldsymbol{Z})\{\beta_a(\boldsymbol{Z}) - \beta(\boldsymbol{Z})\}\Big]\pi_a^{-1} \\
&\quad + E\Big[\{\beta_b(\boldsymbol{Z}) - \beta(\boldsymbol{Z})\}^{\mathrm{T}}\text{var}(\boldsymbol{U} \mid \boldsymbol{Z})\{\beta_b(\boldsymbol{Z}) - \beta(\boldsymbol{Z})\}\Big]\pi_b^{-1} \\
&\quad - E\Big[\{\beta_a(\boldsymbol{Z}) - \beta_b(\boldsymbol{Z})\}^{\mathrm{T}}\,\text{var}(\boldsymbol{U} \mid \boldsymbol{Z})\,\{\beta_a(\boldsymbol{Z}) - \beta_b(\boldsymbol{Z})\}\Big].
\end{aligned}$$

and, hence, $\widehat{\theta}_A$ is asymptotically more efficient than $\widehat{\theta}_B$ unless

$$\begin{aligned}
\beta(\boldsymbol{z}) = \frac{\pi_b\beta_a(\boldsymbol{z}) + \pi_a\beta_b(\boldsymbol{z})}{\pi_a + \pi_b} \quad &\text{and} \\
\{\beta_a(\boldsymbol{z}) - \beta_b(\boldsymbol{z})\}(1 - \pi_a - \pi_b) = 0 \quad &\\
\text{for every } \boldsymbol{z}, &
\end{aligned} \tag{12}$$

in which case $\widehat{\theta}_B$ and $\widehat{\theta}_A$ have the same asymptotic efficiency.

Note that $\widehat{\beta}(\boldsymbol{z})$ used in $\widehat{\theta}_B$ ignores the fact that $\text{cov}(\boldsymbol{U}, Y^{(a)} \mid \boldsymbol{Z} = \boldsymbol{z})$ may depend on treatment a. That is why $\widehat{\theta}_B$ is asymptotically not as efficient as $\widehat{\theta}_A$ in general, and $\sigma_B^2 = \sigma_A^2$ when these covariances are the same for every a and every $\boldsymbol{z}$, i.e., $\beta_1(\boldsymbol{z}) = \cdots = \beta_k(\boldsymbol{z})$ so that (12) holds. If (12) holds, $\widehat{\theta}_B$ may have better finite sample performance than $\widehat{\theta}_A$, although two estimators are asymptotically equivalent. An exceptional case for $\sigma_A^2 = \sigma_B^2$ is when $k = 2$ and $\pi_1 = \pi_2 = 1/2$, in which we even do not need $\beta_a(\boldsymbol{z}) = \beta_b(\boldsymbol{z})$.

In general, $\widehat{\theta}_B$ may be asymptotically less efficient than $\widehat{\theta}_S$, i.e., covariate adjustment with only the main effects may hurt efficiency, a perspective in Freedman (2008) and Lin (2013). For example, there are scenarios in which (11) holds but (12) does not.

Finally, inference about θ can be carried out based on (10) and the availability of consistent estimators of σ_A^2 and σ_B^2. Some model free consistent variance estimators under any covariate-adaptive randomisation schemes are derived in Ye et al. (2020), which are similar to $\widehat{\sigma}_S^2$ in Section 5.2.

6.2. Can covariate-adaptive randomisation boost efficiency?

We now address Question (C) raised in Section 1 and the beginning of this section: Can a test (or an inference procedure) under covariate-adaptive randomisation be more efficient than it is under simple randomisation?

For the types of covariate-adaptive randomisation schemes described in Section 3, the answer is no, assuming that exactly the same test is used under simple randomisation or under covariate-adaptive randomisation without adjusting for conservativeness. This answer is based on the first-order asymptotic property. With a fixed n, the test or inference procedure under covariate-adaptive randomisation may perform slightly better due to the balancedness of treatment assignments.

In our previous discussions, a conventional procedure may be conservative under covariate-adaptive randomisation, and a valid procedure can often be constructed by modifying the conventional procedure. This modified procedure can be more efficient than the conventional procedure, but the comparison is not fair because the modified procedure makes some adjustment typically depending on the covariate $\boldsymbol{Z}$.

Then, what is the advantage of applying covariate-adaptive randomisation? It is applied mainly for balancing treatment assignments across prognostic factors, which may be important for reviewing clinical results and other practical considerations.

There is a stream of developments and results in balancing discrete or continuous covariates and increasing estimation efficiency at the same time (Atkinson, 1982, 1999, 2002; Baldi Antognini & Zagoraiou, 2011; Rosenberger & Sverdlov, 2008; Senn et al., 2010). The approaches are typically model-based

and the gains in efficiency may be from the second-order asymptotics.

Boosting efficiency can also be achieved by adjusting covariates under simple randomisation with less effort compared with applying covariate-adaptive randomisation, which is discussed next.

6.3. Designing versus modelling

Utilising covariate $\boldsymbol{Z}$ in randomisation can be viewed as a kind of designing for better quality of data, although this is not the same as what in the traditional experiment design, because in clinical trials we typically cannot control covariate values of patients. Adjusting for covariates, either model-based or model-assisted, fits into the general framework of modelling. In this section, we address the issue of designing versus modelling.

First, consider inference on the average treatment effect $\theta = E(Y^{(a)} - Y^{(b)})$. If $\boldsymbol{Z}$ is the only covariate, i.e., $\boldsymbol{X} = \boldsymbol{Z}$, then the conclusion is that designing and modelling (adjusting for covariate) can achieve the same efficiency asymptotically. In this case, $\widehat{\theta}_S = \widehat{\theta}_A = \widehat{\theta}_B$ and it has the same asymptotic normal distribution under simple randomisation and under any other covariate-adaptive randomisation satisfying (D1)–(D2). The stratification in (7) serves the purpose of modelling under simple randomisation, but it is essential for obtaining easy inference under covariate-adaptive randomisation including minimisation.

Consider next the situation where $\boldsymbol{X} \neq \boldsymbol{Z}$ and the covariate $\boldsymbol{U}$ as discussed in Section 6.1 together with $\boldsymbol{Z}$ are available for modelling (the entire covariate $\boldsymbol{X}$ may still contain more information than that from $\boldsymbol{U}$ and $\boldsymbol{Z}$). The conclusion is, modelling with $\boldsymbol{Z}$ and $\boldsymbol{U}$ achieves more efficiency than designing with $\boldsymbol{Z}$ only, and is the same as designing with $\boldsymbol{Z}$ plus an additional modelling with $\boldsymbol{U}$ (adjusting for $\boldsymbol{U}$). This directly comes from result (10). Under simple randomisation, $\widehat{\theta}_A$ in Section 6.1 is the estimator after modelling with $\boldsymbol{Z}$ and $\boldsymbol{U}$, in view of the fact that $\boldsymbol{Z}$ is discrete so that stratification is the same as modelling with $\boldsymbol{Z}$, and its limiting variance is σ_A^2 in (10). On the other hand, designing with $\boldsymbol{Z}$ only leads to the estimator $\widehat{\theta}_S$ in (7), which has limiting variance σ_S^2 in (8) regardless of which covariate-adaptive randomisation is applied, and $\sigma_A^2 \leq \sigma_S^2$. Finally, designing with $\boldsymbol{Z}$ plus an additional modelling with $\boldsymbol{U}$ leads to estimator $\widehat{\theta}_A$.

Similar conclusions can be obtained for testing in survival analysis as discussed in Section 4.4. Consider the situation of $\boldsymbol{X} = \boldsymbol{Z}$. Since $\boldsymbol{Z}$ is discrete, the Cox model given by (5) is always correct. Modelling with $\boldsymbol{Z}$ produces the score test under simple randomisation, whereas designing with $\boldsymbol{Z}$ leads to the stratified log-rank test in (6). It is shown in Ye and Shao (2020) that the two tests have the same Pitman's asymptotic efficiency. If $\boldsymbol{X} \neq \boldsymbol{Z}$ and the Cox model (5) with $\boldsymbol{X}$ is correct, then it is shown in Ye and Shao (2020) that the score test under simple randomisation is more efficient than the stratified log-rank test based on designing and stratification with $\boldsymbol{Z}$, in terms of Pitman's asymptotic efficiency. In this case, designing with $\boldsymbol{Z}$ plus an additional modelling leads to the score test. Unlike the case of inference on average treatment effect, however, in survival testing all results for the situation of $\boldsymbol{X} \neq \boldsymbol{Z}$ relies on the correctness of Cox model (5). If model (5) is wrong, then the score test can be less powerful than the unstratified log-rank test.

7. Further research work

We end this review with the following discussion of further research topics in this area.

(1) Although some estimation and inference procedures previously discussed have asymptotic distributions invariant to covariate-adaptive randomisation schemes, it may be still important to study and understand the Type 3 randomisation methods such as the minimisation whose properties are unclear at this stage. In particular, the asymptotic property of $D^{(a)}(\boldsymbol{z})$ defined in (D2). Efforts should be made to establish the joint asymptotic distribution of $D^{(a)}(\boldsymbol{z})$ with $\boldsymbol{z}$ being all levels of $\boldsymbol{Z}$. A different direction is to develop more and better covariate-adaptive randomisation schemes. For example, in Section 5.3 we point out that a Type 1 randomisation scheme may produce more efficient inference procedures than a Type 2 or 3 randomisation scheme. Hu and Hu (2012) modified Pocock and Simon's approach and proposed to use an imbalance measure that is a weighted sum of the overall imbalance, marginal imbalance, and strata imbalance. Some effort should be made to study the implementation of this scheme for practical uses.

(2) To utilise covariates, we considered the model-assisted generalised regression approach for the estimation of average treatment effect and score test under a working Cox model for testing hypotheses in survival analysis. It is interesting to develop other model-assisted approaches to gain efficiency without relying on models.

(3) From result (8), if $\boldsymbol{Z}'$ is another covariate such that the σ-field of $\boldsymbol{Z}'$ contains the σ-field of $\boldsymbol{Z}$, then the $\widehat{\theta}_A$ using $\boldsymbol{Z}'$ in randomisation is asymptotically more efficient than the $\widehat{\theta}_A$ using $\boldsymbol{Z}$ in randomisation. That is, utilising more covariate information in randomisation can increase asymptotic efficiency. On the other hand, using a $\boldsymbol{Z}$ with too many levels may cause sparsity of data. Some guidance on this may be useful for practical users.

(4) The stratification in (6) or (7) uses all levels of $\boldsymbol{Z}$ as strata. In applications, it is possible that some strata

contain very few number of patients or even no patient. Some methods of handling this scenario should be developed to produce asymptotically valid or at least conservative inference procedures, such as combining some strata with small sizes.

(5) The result and discussion on inference about quantile treatment effects are very limited. In survival analysis, due to the presence of censoring, the distribution function estimator in (9) has to be replaced by the Kaplan–Meier product-limit type estimator. Furthermore, how to adjust for covariates has not been considered.

(6) The bootstrap, re-randomisation and permutation methods described in Section 4.3 are promising alternative tools to the approach of asymptotic distribution plus variance estimation for statistical inference. Two issues have to be addressed. The first one is that the re-randomisation and permutation methods are naturally developed for testing. Applying these tools for inference on parameters other than the average treatment effect requires further development. The other issue is what we discussed in the end of Section 4.4, i.e., the development of bootstrap or re-randomisation methods when censoring distributions conditioned on $\boldsymbol{X}$ can be different under the null hypothesis that the survival distributions conditioned on $\boldsymbol{X}$ are identical.

Acknowledgements

Our research was supported by the National Natural Science Foundation of China (11831008) and the U.S. National Science Foundation (DMS-1914411).

Disclosure statement

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

Funding

Our research was supported by the National Natural Science Foundation of China (11831008) and the U.S. National Science Foundation (DMS-1914411).

Notes on contributor

Dr Jun Shao holds a PhD in statistics from the University of Wisconsin-Madison. He is a Professor of Statistics at the University of Wisconsin-Madison. His research interests include variable selection and inference with high dimensional data, sample surveys, and missing data problems.

References

Aickin, M. (2002). Beyond randomization. *The Journal of Alternative and Complementary Medicine*, *8*(6), 765–772. https://doi.org/10.1089/10755530260511775

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, *69*(1), 61–67. https://doi.org/10.1093/biomet/69.1.61

Atkinson, A. C. (1999). Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine*, *18*(14), 1741–1752. https://doi.org/10.1002/(ISSN)1097-0258

Atkinson, A. C. (2002). The comparison of designs for sequential clinical trials with covariate information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(2), 349–373. https://doi.org/10.1111/rssa.2002.165.issue-2

Baldi Antognini, A., & Zagoraiou, M. (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika*, *98*(3), 519–535. https://doi.org/10.1093/biomet/asr021

Baldi Antognini, A., & Zagoraiou, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli*, *21*(2), 881–908. https://doi.org/10.3150/13-BEJ591

Birkett, N. J. (1985). Adaptive allocation in randomized controlled trials. *Controlled Clinical Trials*, *6*(2), 146–155. https://doi.org/10.1016/0197-2456(85)90120-5

Breugom, A. J., van Gijn, W., Muller, E. W., Berglund, Å., Fokstuen, T., Gelderblom, H., Kapiteijn, E., Leer, J. W. H., Marijnen, C. A. M., Martijn, H., Meershoek-Klein Kranenbarg, E., Nagtegaal, I. D., Påhlman, L., Punt, C. J. A., Putter, H., Roodvoets, A. G. H., Rutten, H. J. T., Steup, W. H., Glimelius, B., & C. J. H. van de Velde (2015). Adjuvant chemotherapy for rectal cancer patients treated with preoperative (chemo)radiotherapy and total mesorectal excision: A dutch colorectal cancer group (dccg) randomized phase III trial. *Annals of Oncology*, *26*(4), 696–701.

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, *113*(524), 1784–1796. https://doi.org/10.1080/01621459.2017.1375934

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, *10*(4), 1747–1785. https://doi.org/10.3982/QE1150

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*(3), 615–620. https://doi.org/10.1093/biomet/63.3.615

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), 1715. https://doi.org/10.1186/s12874-019-0787-8

Committee for proprietary medicinal products. (2004). Points to consider on adjustment for baseline covariates. *Statistics in Medicine*, *23*(5), 701–709. https://doi.org/10.1002/(ISSN)1097-0258

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/rssb.2001.63.issue-4

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, *58*(3), 403–417. https://doi.org/10.1093/biomet/58.3.403

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

Fakhry, F., Spronk, S., van der Laan, L., Wever, J. J., Teijink, J. A. W., Hoffmann, W. H., Smits, T. M., van Brussel, J. P., Stultiens, G. N. M., Derom, A., den Hoed, P. T., Ho, G. H., van Dijk, L. C., Verhofstad, N., Orsini, M., van Petersen, A., Woltman, K., Hulst, I., van Sambeek, M. R. H. M., … Hunink, M. G. M. (2015). Endovascular revascularization and supervised exercise for peripheral artery disease and intermittent claudication: A randomized clinical trial. *The Journal of the American Medical Association*, *314*(18), 1936–1944. https://doi.org/10.1001/jama.2015.14851

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*(1), 259–276. https://doi.org/10.1111/ecta.2007.75.issue-1

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. https://doi.org/10.1016/0167-9473(87)90015-6

Freedman, D. A. (2008). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, *2*(1), 176–196. https://doi.org/10.1214/07-AOAS143

Hagino, A., Hamada, C., Yoshimura, I., Ohashi, Y., Sakamoto, J., & Nakazato, H. (2004). Statistical comparison of random allocation methods in cancer clinical trials. *Controlled Clinical Trials*, *25*(6), 572–584. https://doi.org/10.1016/j.cct.2004.08.004

Han, B., Enas, N. H., & McEntegart, D. (2009). Randomization by minimization for unbalanced treatment allocation. *Statistics in Medicine*, *28*(27), 3329–3346. https://doi.org/10.1002/sim.v28:27

Horn, L., Mansfield, A. S., Szczesna, A., Havel, L., Krzakowski, M., Hochmair, M. J., Huemer, F., Losonczy, G., Johnson, M. L., Nishio, M., Reck, M., Mok, T., Lam, S., Shames, D. S., Liu, J., Ding, B., Lopez-Chavez, A., Kabbinavar, F., Lin, W., Sandler, A., & Liu, S. V. (2018). First-line atezolizumab plus chemotherapy in extensive-stage small-cell lung cancer. *New England Journal of Medicine*, *379*(23), 2220–2229. https://doi.org/10.1056/NEJMoa1809064

Hu, Y., & Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics*, *40*(3), 1794–1815. https://doi.org/10.1214/12-AOS983

Hu, F., & Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons.

Hu, F., & Zhang, L. X. (2020). On the theory of covariate-adaptive designs. Working paper.

Hu, F., Zhang, L. X., & He, X. (2009). Efficient randomized-adaptive designs. *The Annals of Statistics*, *37*(5A), 2543–2560. https://doi.org/10.1214/08-AOS655

Jourdain, G., Ngo-Giang-Huong, N., Harrison, L., Decker, L., Khamduang, W., Tierney, C., Salvadori, N., Cressey, T. R., Sirirungsi, W., Achalapong, J., Yuthavisuthi, P., Kanjanavikai, P., Ayudhaya, O. P. N., Siriwachirachai, T., Prommas, S., Sabsanong, P., Limtrakul, A., Varadisai, S., Putiyanun, C., … Chotivanich, N. (2018). Tenofovir versus placebo to prevent perinatal transmission of hepatitis b. *New England Journal of Medicine*, *378*(10), 911–923. https://doi.org/10.1056/NEJMoa1708131

Kaiser, L. D. (2012). Dynamic randomization and a randomization model for clinical trials data. *Statistics in Medicine*, *31*(29), 3858–3873. https://doi.org/10.1002/sim.v31.29

Kalish, L. A., & Begg, C. B. (1985). Treatment allocation methods in clinical trials: A review. *Statistics in Medicine*, *4*(2), 129–144. https://doi.org/10.1002/(ISSN)1097-0258

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

Kuznetsova, O. M., & Johnson, V. P. (2017). Approaches to expanding the two-arm biased coin randomization to unequal allocation while preserving the unconditional allocation ratio. *Statistics in Medicine*, *36*(16), 2483–2498. https://doi.org/10.1002/sim.7290

Lachin, J. M., Matts, J. P., & Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials*, *9*(4), 365–374. https://doi.org/10.1016/0197-2456(88)90049-9

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, *7*(1), 295–318. https://doi.org/10.1214/12-AOAS583

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

Ma, W., Hu, F., & Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*, *110*(510), 669–680. https://doi.org/10.1080/01621459.2014.922469

Ma, W., Qin, Y., Li, Y., & Hu, F. (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, *115*(531), 1488–1497. https://doi.org/10.1080/01621459.2019.1635483

McKeever, T., Mortimer, K., Wilson, A., Walker, S., Brightling, C., Skeggs, A., Pavord, I., Price, D., Duley, L., Thomas, M., Bradshaw, L., Higgins, B., Haydock, R., Mitchell, E., Devereux, G., & Harrison, T. (2018). Quadrupling inhaled glucocorticoid dose to abort asthma exacerbations. *New England Journal of Medicine*, *378*(10), 902–910. https://doi.org/10.1056/NEJMoa1714257

Mehra, M. R., Goldstein, D. J., Uriel, N., Cleveland, J. C., Yuzefpolskaya, M., Salerno, C., Walsh, M. N., Milano, C. A., Patel, C. B., Ewald, G. A., Itoh, A., Dean, D., Krishnamoorthy, A., Cotts, W. G., Tatooles, A. J., U. P. Jorde, Bruckner, B. A., Estep, J. D., Jeevanandam, V., … Naka, Y. (2018). Two-year outcomes with a magnetically levitated cardiac pump in heart failure. *New England Journal of Medicine*, *378*(15), 1386–1395. https://doi.org/10.1056/NEJMoa1800866

Myles, P. S., Bellomo, R., Corcoran, T., Forbes, A., Peyton, P., Story, D., Christophi, C., Leslie, K., McGuinness, S., Parke, R., Serpell, J., Chan, M. T. V., Painter, T., McCluskey, S., Minto, G., & Wallace, S. (2018). Restrictive versus liberal fluid therapy for major abdominal surgery. *New England Journal of Medicine*, *378*(24), 2263–2274. https://doi.org/10.1056/NEJMoa1801601

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

Ramirez, P. T., Frumovitz, M., Pareja, R., Lopez, A., Vieira, M., Ribeiro, R., Buda, A., Yan, X., Shuzhong, Y., Chetty, N., Isla, D., Tamura, M., Zhu, T., Robledo, K. P., Gebski, V., Asher, R., Behan, V., Nicklin, J. L., Coleman, R. L., & Obermair, A. (2018). Minimally invasive versus abdominal radical hysterectomy for cervical cancer.

New England Journal of Medicine, *379*(20), 1895–1904. https://doi.org/10.1056/NEJMoa1806395
Rosenberger, W. F., & Lachin, J. M. (2015). *Randomization in clinical trials: Theory and practice*: Wiley.
Rosenberger, W. F., & Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*, *23*(3), 404–419. https://doi.org/10.1214/08-STS269
Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.
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
Senn, S., Anisimov, V. V., & Fedorov, V. V. (2010). Comparisons of minimization and Atkinson's algorithm. *Statistics in Medicine*, *29*(7–8), 721–730. https://doi.org/10.1002/sim.3763
Shao, J. (2003). *Mathematical statistics* (2nd ed.). Springer.
Shao, J., & Wang, S. (2014). Efficiency of model-assisted regression estimators in sample surveys. *Statistica Sinica*, *24*(1), 395–414.
Shao, J., & Yu, X. (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics*, *69*(4), 960–969. https://doi.org/10.1111/biom.12062
Shao, J., Yu, X., & Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, *97*(2), 347–360. https://doi.org/10.1093/biomet/asq014
Simon, R., & Simon, N. R. (2011). Using randomization tests to preserve type I error with response-adaptive and covariate-adaptive randomization. *Statistics & Probability Letters*, *81*(7), 767–772. https://doi.org/10.1016/j.spl.2010.12.018
Stott, D. J., Rodondi, N., Kearney, P. M., Ford, I., Westendorp, R. G., S. P. Mooijaart, Sattar, N., Aubert, C. E., Aujesky, D., Bauer, D. C., Baumgartner, C., Blum, M. R., Browne, J. P., Byrne, S., Collet, T. -H., Dekkers, O. M., Puy, R. S. D., Ellis, G., … Gussekloo, J. (2017). Thyroid hormone therapy for older adults with subclinical hypothyroidism. *New England Journal of Medicine*, *376*(26), 2534–2544.
Sun, J.-M., Lee, K. H., Kim, B.-S., Kim, H.-G., Min, Y. J., Yi, S. Y., Yun, H. J., Jung, S.-H., Lee, S.-H., Ahn, J. S., Park, K., & Ahn, M.-J. (2018). Pazopanib maintenance after first-line etoposide and platinum chemotherapy in patients with extensive disease small-cell lung cancer: A multicentre, randomised, placebo-controlled phase II study (kcsg-lu12-07). *British Journal of Cancer*, *118*(5), 648–653.
Ta, T., Shao, J., Li, Q., & Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statistica Sinica*, *30*(3), 1135–1154.
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
van der Ploeg, A. T., Clemens, P. R., Corzo, D., Escolar, D. M., Florence, J., Groeneveld, G. J., Herson, S., Kishnani, P. S., Laforet, P., Lake, S. L., Lange, D. J., R. T. Leshner, Mayhew, J. E., Morgan, C., Nozaki, K., Park, D. J., Pestronk, A., Rosenbloom, B., Skrinar, A., … Zivkovic, S. A. (2010). A randomized study of alglucosidase alfa in late-onset pompe's disease. *New England Journal of Medicine*, *362*(15), 1396–1406. https://doi.org/10.1056/NEJMoa0909859
Wei, L. J. (1977). A class of designs for sequential clinical trials. *Journal of the American Statistical Association*, *72*(358), 382–386. https://doi.org/10.1080/01621459.1977.10481005
Wei, L. J. (1978a). The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, *6*(1), 92–100. https://doi.org/10.1214/aos/1176344068
Wei, L. J. (1978b). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, *73*(363), 559–563. https://doi.org/10.1080/01621459.1978.10480054
Weir, C. J., & Lees, K. R. (2003). Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in Medicine*, *22*(5), 705–726. https://doi.org/10.1002/(ISSN)1097-0258
Xu, Z., Proschan, M., & Lee, S. (2016). Validity and power considerations on hypothesis testing under minimization. *Statistics in Medicine*, *35*(14), 2315–2327. https://doi.org/10.1002/sim.v35.14
Ye, T. (2018). Testing hypotheses under covariate-adaptive randomisation and additive models. *Statistical Theory and Related Fields*, *2*(1), 96–101. https://doi.org/10.1080/24754269.2018.1477005
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.v82.5
Ye, T., Yi, Y., & Shao, J. (2020). Inference on average treatment effect under minimization and other covariate-adaptive randomization methods, arXiv:2007.09576.
Zannad, F., Anker, S. D., Byra, W. M., Cleland, J. G. F., Fu, M., Gheorghiade, M., Lam, C. S. P., Mehra, M. R., Neaton, J. D., Nessel, C. C., Spiro, T. E., D. J. van Veldhuisen, & Greenberg, B. (2018). Rivaroxaban in patients with heart failure, sinus rhythm, and coronary disease. *New England Journal of Medicine*, *379*(14), 1332–1342. https://doi.org/10.1056/NEJMoa1808848
Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Clinical Epidemiology*, *27*(7), 365–375.
Zhang, L. X., Hu, F., Cheung, S. H., & Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, *35*(3), 1166–1182. https://doi.org/10.1214/009053606000001424
Zhang, Y., Wang, L., Yu, M., & Shao, J. (2020). Quantile treatment effect estimation with dimension reduction. *Statistical Theory and Related Fields*, *4*(2), 202–213. https://doi.org/10.1080/24754269.2019.1696645
Zhao, W., & Ramakrishnan, V. (2016). Generalization of Wei's urn design to unequal allocations in sequential clinical trials. *Contemporary Clinical Trials Communications*, *2*, 75–79. https://doi.org/10.1016/j.conctc.2015.12.007
Zhong, B., & Kim, L. (2008). Adaptive randomization, the preferred randomization in clinical trials. In *Proceedings of the American statistical association* (pp. 3460–3467).