SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(6) 1715–1727

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280219846146
journals.sagepub.com/home/smm
SAGE

Design and analysis of stratified clinical trials in the presence of bias

Ralf-Dieter Hilgers,[1] Martin Manolov,[1] Nicole Heussen[1,2] and William F Rosenberger[3]

Abstract

Background: Among various design aspects, the choice of randomization procedure have to be agreed on, when planning a clinical trial stratified by center. The aim of the paper is to present a methodological approach to evaluate whether a randomization procedure mitigates the impact of bias on the test decision in clinical trial stratified by center.
Methods: We use the weighted *t* test to analyze the data from a clinical trial stratified by center with a two-arm parallel group design, an intended 1:1 allocation ratio, aiming to prove a superiority hypothesis with a continuous normal endpoint without interim analysis and no adaptation in the randomization process. The derivation is based on the weighted *t* test under misclassification, i.e. ignoring bias. An additive bias model combing selection bias and time-trend bias is linked to different stratified randomization procedures.
Results: Various aspects to formulate stratified versions of randomization procedures are discussed. A formula for sample size calculation of the weighted *t* test is derived and used to specify the tolerated imbalance allowed by some randomization procedures. The distribution of the weighted *t* test under misclassification is deduced, taking the sequence of patient allocation to treatment, i.e. the randomization sequence into account. An additive bias model combining selection bias and time-trend bias at strata level linked to the applied randomization sequence is proposed. With these before mentioned components, the potential impact of bias on the type one error probability depending on the selected randomization sequence and thus the randomization procedure is formally derived and exemplarily calculated within a numerical evaluation study.
Conclusion: The proposed biasing policy and test distribution are necessary to conduct an evaluation of the comparative performance of (stratified) randomization procedure in multi-center clinical trials with a two-arm parallel group design. It enables the choice of the best practice procedure. The evaluation stimulates the discussion about the level of evidence resulting in those kind of clinical trials.

Keywords
Multi-center clinical trial, weighted *t* test, sample size, stratified randomization, type I error probability, selection bias, time-trend bias

1 Introduction

Large clinical trials often stratify the randomization on a small collection of covariates that may introduce heterogeneity into the patient stream. An important covariable in multi-center trials is often the clinical center, as different study personnel, clinical settings, and patient populations may result in differential study outcomes.[1] A stratified population-based analysis can be performed with or without stratification in the design. Less is known about the impact of stratification when there is a bias in the clinical trial. In this paper, we explore this issue both for selection bias and chronological bias, and we demonstrate the impact of these analyses on a weighted stratified analysis. In so doing, we explore the role of specific stratified randomization procedures (RPs) and how certain procedures may mitigate the effects of bias. The recognition of the role of RPs in mitigating bias has been explored

[1]Department of Medical Statistics, RWTH Aachen University, Aachen, Germany
[2]Department of Biostatistics, Sigmund Freud University, Vienna, Austria
[3]Department of Statistics, George Mason University, Fairfax, VA, USA

Corresponding author:
Ralf-Dieter Hilgers, Department of Medical Statistics, RWTH Aachen University, Pauwelsstr 30, D52074 Aachen, Germany.
Email: rhilgers@ukaachen.de

in prior research for unstratified trials.[2–6] But because stratification into K strata creates K different independent randomized clinical trials, and a stratified test combines K independent tests, the impact of bias can be more pronounced.

The paper is organized as follows: In Section 2, we describe different stratified RPs and discuss aspects to formulate stratified versions of RPs. In Section 3, we derive a formulation of Fleiss[1] stratified test statistic preserving the allocation sequence and derive the distribution of the test statistic taking bias into account but ignoring bias in the analysis and mention some sample size considerations for the stratified test. In Section 4, we specify the bias model in the form of an additive combination of strata-specific selection bias and strata-specific time-trend bias linked to stratified allocation sequence. The criterion introduced in Section 5 is used to summarize the impact of the allocation sequence-specific bias on the type I error probability over the range of all sequences induced by a specific RP. Consequently, an assessment of different RPs is enabled which guides the choice of an RPs for application in a particular clinical trial setting. The methodology is applied to some-specific scenarios in Section 6 to illustrate the effects. We discuss the findings in Section 7 and draw conclusions in Section 8.

2 Stratified RPs

RPs for clinical trials for two treatments are well described in literature.[2] In principle, any RP used for two-treatment clinical trials can be employed within strata in a stratified randomization. A comprehensive review is given in Rosenberger.[2] Complete randomization in which patients are assigned to treatments with probability 1/2 is rarely used in stratified clinical trials. Rather, some form of restricted randomization is employed in an effort to balance treatments within strata. Hilgers[6] categorized restricted RPs that force balance in probability, force balance using a maximal tolerated imbalance, or force terminal balance. A selective list of restricted RPs is given as follows[2]:

- *Efron's biased coin design* (EBC(p)), which consists of flipping a biased coin with probability $p \geq 0.5$ in favor of the treatment which has been allocated less frequently and a fair coin in case of equal numbers of treatment assignments,[7]
- *Big stick design* (BSD(a)), which can be implemented via complete randomization with a forced deterministic assignment when a maximal tolerated imbalance a is reached during the enrollment,[8]
- *Random allocation rule* (RAR), which assigns half the patients to E and C randomly,[9]
- *Permuted block randomization* (PBR(b)) with block size b uses RAR within blocks of b patients, for b even,[10]
- *Maximal procedure* (MP(a)) which uses the allocation sequences of RAR by additionally imposing a maximal tolerated imbalance (a) and assigning equal probability to all such sequences.[11]

Note that EBC(p) may be classified as a restricted RP forcing balance in probability. BSD(a) forces balance by maximal tolerated imbalance a during the allocation process but does not force terminal balance. Restricted RPs with a maximal tolerable imbalance and terminal balance are PBR(b) and MP(a).

The International Council of Harmonization stated in the E9 recommendation (ICH E9)

> It is advisable to have a separate random scheme for each centre, i.e., to stratify by centre or to allocate several whole blocks to each centre.

The European Medicines Agency "Guideline on Clinical Trials in Small Populations" recommends stratified randomization to improve power. Using permuted blocks within each stratum is the most popular method of stratified randomization, and this is often called the *stratified block design*. Blocks can be selected with a fixed size or with variable sizes. However, blocking is not the only method to use within strata. The ICH E9[12] guidelines also state that "different trial designs will require different procedures for generating randomization schedules." We now define stratified randomization more formally.

Consider the allocation $z_{ji} \in \{0, 1\}$ of patients $i = 1, \ldots, n_j$ either to the treatment E if $z_{ji} = 1$ or C if $z_{ji} = 0$ in stratum j. An RP is implemented by assigning probabilities $P(\mathbf{Z}_j = \mathbf{z}_j | \mathbf{z}_j \in \{0, 1\}^{n_j})$ to the possible allocations $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jn_j})$ in stratum j. A *stratified randomization* is implemented by creating independent randomization lists $\mathbf{z}_j \in \{0, 1\}^{n_j}$ for each stratum $1 \leq j \leq K$. Denote the possible allocations by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_K) \in \times_{j=1}^{K} \{0, 1\}^{n_j} = \{0, 1\}^N$ with $N = \sum_{j=1}^{K} n_j$. Then a stratified version of an RP is implemented by assigning probabilities $P(\mathbf{Z}=\mathbf{z}|\mathbf{z}\in\{0,1\}^N):=\prod_{j=1}^{K}P(\mathbf{Z}_j=\mathbf{z}_j|\mathbf{z}_j\in\{0,1\}^{n_j})$ to the possible allocations $\mathbf{z}\in\{0,1\}^N$. Of course, when implementing complete randomization, the stratified and unstratified RPs result in the same set if randomization sequences with the same probabilities because assignments are independent and equiprobable, i.e.

$$P(\mathbf{Z}=\mathbf{z}|\mathbf{z}\in\{0,1\}^N)=\prod_{j=1}^{K}P(\mathbf{Z}_j=\mathbf{z}_j|\mathbf{z}_j\in\{0,1\}^{n_j})=\prod_{j=1}^{K}\frac{1}{2^{n_j}}=\frac{1}{2^N}$$

However, when implementing a stratified restricted RP, this observation generally does not hold and some further definitions are necessary. Even in the very simple RAR, the set of possible randomization sequences is reduced considerable and the probability for the stratified allocation sequence becomes

$$P(\mathbf{Z}=\mathbf{z}|\mathbf{z}\in\{0,1\}^N)=\prod_{j=1}^{K}P(\mathbf{Z}_j=\mathbf{z}_j|\mathbf{z}_j\in\{0,1\}^{n_j})=\prod_{j=1}^{K}\binom{n_j}{n_j/2}^{-1}\neq\binom{N}{N/2}^{-1}$$

Another important aspect concerns the ‘‘balancing behavior’’ of restricted RPs. The term restricted refers to the fact that conditions on the randomization process are introduced to control the potential imbalance in the frequency of treatment allocations. Let $s, 1\le s\le N,$ denotes the patient's number preserving the appearance of patients in the trial so that $s=1$ denotes the first patient and $s=N$ the last enrolled patient. $n_{jE}(s)$ and $n_{jC}(s)$ denote the number of patients allocated to treatments E and C in stratum j until a total of s patients are recruited in the trial so far. Then, the imbalance in the number of allocations to treatment E and C in stratum j until a total of s patients are recruited is measured by

$$d_j(s)=n_{jE}(s)-n_{jC}(s) \tag{1}$$

Three definitions of imbalance are used in the following:

(1) An RP shows *overall final balance*, if $d(N)\overset{\text{def}}{=}\sum_{j=1}^{K}d_j(N)=0$
(2) An RP controls the *final balance within strata*, if $d_j(N)=0$ for $1\le j\le K$
(3) An RP controls the *maximal tolerated imbalance*, if $-a\le d_j(s)\le a$ for all $1\le j\le K, 1\le s\le N$.

Of course controlling the overall final balance within strata does not imply to control final balance within stratum, i.e. $d_j(s)=0$. Simply controlling the overall final balance may result in one stratum assigning patients only to E and another stratum assigning the same number of patients to C only, a case which invalidates the estimation of treatment difference within strata, presumably one issue in the ICH guidance. On the other hand, final balance within strata ($d_j(N)=0$) implies overall final balance $d(N)=0$. In the following, we deal with stratified RPs and derive additional restrictions for meaningful definitions. With RAR, stratification and consequently final balance within strata require even samples sizes within each stratum and two treatment arms. The requirement of final balance within strata implies in the case of the stratified block design that the block sizes are divisors of the stratum sample sizes $n_j, 1\le j\le K$. Note that in stratified trials with a larger number of centers, usually smaller sample sizes in centers occur and thus final balance within strata forces the block sizes to be small, which will increase the potential for selection bias. Of course, center-specific block sizes are possible but rather uncommon. We will consider common block sizes in the following.

It should also be noted that the stratified RAR procedure in general cannot be considered as an unstratified PBR with block sizes $n_j, 1\le j\le K,$ because enrollment of patients in the trial is parallel in strata, so that in general $d\left(\sum_{j=1}^{k}n_j\right)\neq 0, 1\le k\le K$.

Similar problems arise, when controlling the maximal tolerated imbalance with margin a, which results in an upper overall bound of $\sum_{j=1}^{K}|d_j(s)|\le Ka$ for all $1\le s\le N$. Thus, controlling the maximal tolerated imbalance across strata could be accomplished by having a different imbalance level in each stratum. Two very straightforward simple settings are uniform spread $|d_j(s)|\le a/K, 1\le j\le K$ for all $1\le s\le N$ resulting $\sum_{j=1}^{K}|d_j(s)|\le\sum_{j=1}^{K}a/K=a$ for all $1\le s\le N$ or proportional spread with $|d_j(s)|\le an_j/N, 1\le j\le K$ for all $1\le s\le N$ resulting $\sum_{j=1}^{K}|d_j(s)|\le\sum_{j=1}^{K}an_j/N=a$ for all $1\le s\le N$. Hilgers[6] suggests defining a in relation to loss of power, and this implies stratum-specific maximal tolerated imbalance according to the rules above.

3 Stratified analysis

As mentioned in Section 1, a stratified randomization requires a stratified analysis, although a stratified analysis can be performed whether or not the randomization was stratified. In this section, we examine the distributional properties of a test statistic introduced by Fleiss[1] (page 268, formulas 1 and 2) based on a weighted t statistic for the analysis of stratified clinical trials. While we do not consider randomization tests in this paper, clearly randomization-based inference is an attractive alternative, see Rosenberger.[2] The reason for using a parametric t test is that it facilitates our goal of determining the effect of bias on inference, since we can derive the distribution of the test statistic under various forms of bias. In particular, in this section, we derive the non-centrality parameter for the distribution of the test statistic under alternative hypotheses and comment on how it can be used for sample size considerations. In the sequel, we are interested in the role of the RP in the analysis of stratified trials. Because Fleiss wrote specifically about centers rather than strata, we use both interchangeably; it should be clear that stratification can be done on variables other than center however.

We will consider a two-arm parallel group clinical trial stratified by K centers with no interim analysis. The response to the treatments E and C respectively is measured with the continuous normally distributed endpoint $y_{ji}, 1 \leq i \leq n_j = n_{jE} + n_{jC}, 1 \leq j \leq K,$ on n_{jE} patients in the experimental group (E) and n_{jC} patients in the control group (C) in centers j. The total sample size is denoted by $N = \sum_{j=1}^{K} n_j$.

We use the *allocation sequence notation* of the statistical model assuming no treatment by center interaction by

$$y_{ji} = \mu_E Z_{ji} + \mu_C(1 - Z_{ji}) + \tau_{ji} + \epsilon_{ji} \tag{2}$$

where $\epsilon_{ji}N(0, \sigma^2), 1 \leq i \leq n_j = n_{jE} + n_{jC}, 1 \leq j \leq K$. The expected treatment effects under E and C are denoted by μ_E and μ_C, respectively. The Z_{ji} denotes the allocation sequence indicator with $Z_{ji} = 1$ if patient i in center j is allocated to treatment E and $Z_{ji} = 0$ if patient i in center j is allocated to treatment C. Here and in what follows the notations $n_{jE} = \sum_{i=1}^{n_j} Z_{ji}$ and $n_{jC} = \sum_{i=1}^{n_j} (1 - Z_{ji})$ are used. Furthermore, τ_{ji} denotes the fixed "bias" effect acting on the response of patient i in center j. Without loss of generality, we assume $\tau_{ji} > 0$.

Fleiss's statistic to test the hypothesis ($H_0 : \mu_E = \mu_C$) of no treatment effect across centers becomes

$$t = \frac{\sum_{j=1}^{K} w_j D_j}{s_p \sqrt{\sum_{j=1}^{K} w_j^2 / w_j^*}} = \frac{\left(\sum_{j=1}^{K} w_j D_j\right) \sigma^{-1} \sqrt{\sum_{j=1}^{K} w_j^2 / w_j^*}^{\,-1}}{s_p / \sigma} \tag{3}$$

where $D_j = \tilde{y}_{jE} - \tilde{y}_{jC}$ are the mean treatment differences with $\tilde{y}_{jE} = \frac{1}{n_{jE}} \sum_{i=1}^{n_j} y_{ji} Z_{ji}$ and $\tilde{y}_{jC} = \frac{1}{n_{jC}} \sum_{i=1}^{n_j} y_{ji}(1 - Z_{ji})$. Furthermore, w_j are weights associated with center j, $w_j^* = \frac{n_{jE} \times n_{jC}}{n_{jE} + n_{jC}}$ and s_p is the pooled variance given by

$$s_p^2 = \left(\sum_{j=1}^{K} \sum_{\ell = E,C} (n_{j\ell} - 1) s_{j\ell}^2 \right) \Bigg/ \left(\sum_{j=1}^{K} \sum_{\ell = E,C} (n_{j\ell} - 1) \right)$$

Here $s_{j\ell}^2$ denotes the variance of treatment group ℓ in center j. To derive the distribution of equation (3) under model (2), the distributions of the numerator as well denominator must be calculated. Note that the variance is given by

$$\begin{aligned} Var\left(\sum w_j D_j\right) &= \sum_{j=1}^{K} w_j^2 \left(\frac{1}{n_{jE}^2} \sum_{i=1}^{n_j} Var(y_{ji}) Z_{ji} + \frac{1}{n_{jC}^2} \sum_{i=1}^{n_j} Var(y_{ji})(1 - Z_{ji}) \right) \\ &= \sum_{j=1}^{K} w_j^2 \left(\frac{\sigma^2}{n_{jE}} + \frac{\sigma^2}{n_{jC}} \right) = \sigma^2 \sum_{j=1}^{K} w_j^2 \, \frac{n_{jE} + n_{jC}}{n_{jE} \times n_{jC}} = \sigma^2 \sum_{j=1}^{K} w_j^2 / w_j^* \end{aligned}$$

so that the numerator of equation (3) has variance 1. Of course, D_j is normally distributed via the distribution of y_{ij} and thus the expectation of the denominator equals

$$\delta(\mathbf{Z}) = E\left(\frac{\left(\sum_{j=1}^{K} w_j D_j\right)}{\sigma \sqrt{\sum_{j=1}^{K} w_j^2 / w_j^*}} \right)$$

$$= \left(\sigma\sqrt{\sum_{j=1}^{K} w_j^2/w_j^*}\right)^{-1} E\left(\sum_{j=1}^{K} w_j\left(\frac{1}{n_{jE}}\sum_{i=1}^{n_j} y_{ji}Z_{ji} - \frac{1}{n_{jC}}\sum_{i=1}^{n_j} y_{ji}(1-Z_{ji})\right)\right)$$

$$= \left(\sigma\sqrt{\sum_{j=1}^{K} w_j^2/w_j^*}\right)^{-1} \left((\mu_E-\mu_C)\sum_{j=1}^{K} w_j + \sum_{j=1}^{K} w_j\big(\tilde{\tau}_{jE}-\tilde{\tau}_{jC}\big)\right)$$

where $\mathbf{Z} = (Z_{11}, \ldots, Z_{1n_1}, \ldots, Z_{K1}, \ldots, Z_{Kn_K})^t$ is the observed allocation vector, $\tilde{\tau}_{jE} = \frac{1}{n_{jE}}\sum_{i=1}^{n_j} \tau_{jE}Z_{ji}$ and $\tilde{\tau}_{jC} = \frac{1}{n_{jC}}\sum_{i=1}^{n_j} \tau_{jC}(1-Z_{ji})$. In summary, the numerator is i.i.d. normally distributed with expectation $\delta(\mathbf{Z})$ and variance 1.

Next, we calculate the distribution of the denominator s_p/σ using the allocation sequence notation

$$\left(\sum_{j=1}^{K}\sum_{\ell=E,C}(n_{j\ell}-1)\right)\frac{s_p^2}{\sigma^2} = \sum_{j=1}^{K}\sum_{\ell=E,C}(n_{j\ell}-1)\frac{s_{j\ell}^2}{\sigma^2} = \sum_{j=1}^{K}\left((n_{jE}-1)\frac{s_{jE}^2}{\sigma^2} + (n_{jC}-1)\frac{s_{jC}^2}{\sigma^2}\right)$$
$$= \sum_{j=1}^{K}\left(\sum_{i=1}^{n_j} Z_{ji}\Big(\frac{y_{ji}}{\sigma}-\frac{\tilde{y}_{jE}}{\sigma}\Big)^2 + \sum_{i=1}^{n_j}(1-Z_{ji})\Big(\frac{y_{ji}}{\sigma}-\frac{\tilde{y}_{jC}}{\sigma}\Big)^2\right)$$

Note that $Var(y_{ji}/\sigma) = 1$ and $E(y_{ji}/\sigma) = (\mu_E + \tau_{ji})/\sigma$ for $Z_{ji} = 1$ and or $E\big(y_{ji}/\sigma\big) = \big(\mu_C + \tau_{ji}\big)/\sigma$ for $Z_{ji} = 0$ are i.i.d. normally distributed for all $1 \le i \le n_j$ and $1 \le j \le K$. Following the arguments in Johnson and Kotz,[13] the $\sum_{i=1}^{n_j} Z_{ji}(y_{ji} - \tilde{y}_{jE})^2/\sigma^2$ for group E, i.e. $Z_{ij} = 1$ and the $\sum_{i=1}^{n_j}(1-Z_{ji})(y_{ji}-\tilde{y}_{jC})^2/\sigma^2$ for group C, i.e. $Z_{ij} = 0$ are χ^2 distributed with $n_{jE} - 1$ and $n_{jC} - 1$ degrees of freedom respectively and non-centrality parameters

$$\sum_{i=1}^{n_j}\frac{Z_{ji}}{\sigma^2}\left(\mu_E + \tau_{ji} - \frac{1}{n_{jE}}\sum_{i=1}^{n_{jE}}(\mu_E + \tau_{jE})\right)^2 = \sum_{i=1}^{n_j}\frac{Z_{ji}}{\sigma^2}(\tau_{ji}-\tilde{\tau}_{jE})^2$$
$$\sum_{i=1}^{n_j}\frac{(1-Z_{ji})}{\sigma^2}\left(\mu_C + \tau_{ji} - \frac{1}{n_{jC}}\sum_{i=1}^{n_{jC}}(\mu_C + \tau_{jC})\right)^2 = \sum_{i=1}^{n_j}\frac{(1-Z_{ji})}{\sigma^2}(\tau_{ji}-\tilde{\tau}_{jC})^2$$

Applying that the sum of independent non-central $\chi^2_{\nu_j}(\lambda_j)$ distributions is non-central χ^2 with $\sum \nu_j$ degrees of freedom and non-centrality parameter $\sum \lambda_j$, it follows that the distribution of $\Big(\sum_{j=1}^{K}\sum_{\ell=E,C}(n_{j\ell}-1)\Big)s_p^2/\sigma^2$ is non-central χ^2 with non-centrality parameter

$$\lambda(\mathbf{Z}) = \frac{1}{\sigma^2}\sum_{j=1}^{K}\left(\sum_{i=1}^{n_j} Z_{ji}(\tau_{ji}-\tilde{\tau}_{jE})^2 + \sum_{i=1}^{n_j}(1-Z_{ji})(\tau_{ji}-\tilde{\tau}_{jC})^2\right)$$
$$= \frac{1}{\sigma^2}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\tau_{ji}^2 - \sum_{j=1}^{K} n_{jE}\tilde{\tau}_{jE}^2 - \sum_{j=1}^{K} n_{jC}\tilde{\tau}_{jC}^2$$

and

$$df = \sum_{j=1}^{K}\sum_{\ell=E,C}(n_{j\ell}-1) = N - 2K \tag{4}$$

degrees of freedom. Finally, we have to show the independence of the numerator

$$\sum_{j=1}^{K} w_j D_j = \sum_{j=1}^{K} w_j\big(\tilde{y}_{jE}-\tilde{y}_{jC}\big) = \sum_{j=1}^{K} w_j\left(\frac{1}{n_{jE}}\sum_{i=1}^{n_j} y_{ji}Z_{ji} - \frac{1}{n_{jC}}\sum_{i=1}^{n_j} y_{ji}(1-Z_{ji})\right)$$

and denominator

$$\left(\sum_{j=1}^{K}\sum_{\ell=E,C}(n_{j\ell}-1)\right)s_p^2 = \sum_{j=1}^{K}\left(\sum_{i=1}^{n_j} Z_{ji}\big(y_{ji}-\tilde{y}_{jE}\big)^2 + \sum_{i=1}^{n_j}(1-Z_{ji})\big(y_{ji}-\tilde{y}_{jC}\big)^2\right)$$

as random variables. Here, Theorem 3 of Searle[14] is used, stating that two random variables that can be expressed as $\mathbf{x}^t\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$, where $\mathbf{x} \sim N(\mu, \mathbf{V})$ is independent, if $\mathbf{BVA} = 0$. First, note that $\mathbf{V} = \sigma^2\mathbf{I}$ holds in our case.

For enabling the matrix notation of the above expressions, a usual design matrix $\mathbf{X}$ can be defined which includes two columns for the allocation indicator variables and N rows. Rearrangement of the design matrix by center and treatment group so that the first n_{1E} observations belong to treatment E and the preceding n_{1C} observations belong to C in center 1 and so on can be implemented by a suitable permutation matrix $\mathbf{P}$. This simplifies the matrix notation of the above numerator and denominator in terms of $\mathbf{B}$ and $\mathbf{A}$ by reshuffling the allocation sequence $\mathbf{Z} = (Z_{11}, \ldots, Z_{kn_k})^t$ using a suitable permutation matrix $\mathbf{P}$. This permutation matrix does not affect the matrix equation. Furthermore, note that it is sufficient to show the matrix equation for a particular center j because of the block structure implied by the independent observations in different centers. With this reshuffling, the notation for center j corresponding to Theorem 3 is

$$\mathbf{B}_j = w_j\left(\frac{1}{n_{jE}}\mathbf{1}^t_{n_{jE}}, -\frac{1}{n_{jC}}\mathbf{1}^t_{n_{jC}}\right)^t$$

and with $\mathbf{H}_{ij} = \mathbf{I}_{n_{ij}} - \frac{1}{n_{ij}}\mathbf{1}_{n_{ij}\times n_{ij}}$ the matrix

$$\mathbf{A}_j = (\mathbf{H}_{n_{jE}}, \mathbf{H}_{n_{jC}}) = \left(\mathbf{I}_{n_{jE}} - \frac{1}{n_{jE}}\mathbf{1}_{n_{jE}\times n_{jE}}, \mathbf{I}_{n_{jC}} - \frac{1}{n_{jC}}\mathbf{1}_{n_{jC}\times n_{jC}}\right)$$

so that $\sigma^2\mathbf{B}_j\mathbf{I}_{n_{jE}+n_{jC}}\mathbf{A}_j = \mathbf{0}$ for all $1 \le j \le K$, which shows the independence. In summary, the distribution of the statistic in equation (3) is doubly non-central t,[13] with non-centrality parameter

$$\begin{aligned}\delta(\mathbf{Z}) &= \left(\sigma\sqrt{\sum_{j=1}^{K} w_j^2/w_j^*}\right)^{-1}\left((\mu_E - \mu_C)\sum_{j=1}^{K} w_j + \sum_{j=1}^{K} w_j(\tilde{\tau}_{jE} - \tilde{\tau}_{jC})\right)\\ \lambda(\mathbf{Z}) &= \frac{1}{\sigma^2}\left(\sum_{j=1}^{K}\sum_{i=1}^{n_j}\tau_{ji}^2 - \sum_{j=1}^{K} n_{jE}\tilde{\tau}_{jE}^2 - \sum_{j=1}^{K} n_{jC}\tilde{\tau}_{jC}^2\right)\end{aligned} \tag{5}$$

In the case sampling is "stratified" by center and the objective is to estimate the overall treatment effect accounting for center, Fleiss[1] proposed the weights $w_j = w_j^* = \frac{n_{jE}\times n_{jC}}{n_{jE}+n_{jC}}$ resulting in the test statistic (3)

$$t = \frac{\sum_{j=1}^{K} w_j^* D_j}{s_p\sqrt{\sum_{j=1}^{K} w_j^*}} = \frac{\sum_{j=1}^{K}\frac{n_{jE}\times n_{jC}}{n_{jE}+n_{jC}}D_j}{s_p\sqrt{\sum_{j=1}^{K}\frac{n_{jE}\times n_{jC}}{n_{jE}+n_{jC}}}} \tag{6}$$

Of course, equation (5) implies that $\delta(\mathbf{Z})$ depends on the weights only and becomes

$$\delta(\mathbf{z}) = \left(\sigma\sqrt{\sum_{j=1}^{K} w_j^*}\right)^{-1}\left((\mu_E - \mu_C)\sum_{j=1}^{K} w_j^* + \sum_{j=1}^{K} w_j^*(\tilde{\tau}_{jE} - \tilde{\tau}_{jC})\right) \tag{7}$$

In the case sampling is "stratified" by center and the objective is to estimate the overall treatment effect, Fleiss[1] proposed the weights $w_j = 1$ so that equation (3) becomes

$$t = \frac{\sum_{j=1}^{K} D_j}{s_p\sqrt{\sum_{j=1}^{K} 1/w_j^*}}$$

whereas the first non-centrality parameter $\delta(\mathbf{z})$ equals

$$\delta(\mathbf{z}) = \left(\sigma\sqrt{\sum_{j=1}^{K} 1/w_j^*}\right)^{-1}\left(K\cdot(\mu_E - \mu_C) + \sum_{j=1}^{K}(\tilde{\tau}_{jE} - \tilde{\tau}_{jC})\right)$$

Weighting centers in the absence and presence of center-by-treatment interaction has discussed in detail by other authors.[15]

3.1 Sample size considerations

We now briefly discuss the aspects of the sample size and power calculation using the weighted t test statistic. Details can be found in the Supplementary Material Section S1. The results will be used in our numerical evaluation study.

Assuming no bias $\tau_{ji} = 0$ in model (2), the sample size to prove the hypothesis $H_0 : \mu_E = \mu_C$ vs. $\tilde{H}_1 : \mu_E - \mu_C = \Delta$ with the weighted t test (equation (3)) is given by

$$\frac{\left(\sum_{j=1}^{K} w_j\right)^2}{\sum_{j=1}^{K} \frac{w_j^2}{w_j^*}} = \frac{\sigma^2}{\Delta^2}(t_{N-2K}(1-\beta) + t_{N-2K}(1-\alpha/2))^2 \tag{8}$$

The derivation assumed homogeneous variances in all groups and centers. Using the optimal weights of Fleiss,[1] i.e. $w_j = w_j^* = \frac{n_{jE} \times n_{jC}}{n_{jE} + n_{jC}}$, equation (8) simplifies to

$$\sum_{j=1}^{K} \frac{n_{jE} \times n_{jC}}{n_{jE} + n_{jC}} = \frac{\sigma^2}{\Delta^2}(t_{N-2K}(1-\beta) + t_{N-2K}(1-\alpha/2))^2$$

which in case of a balanced allocation ratio of $r \cdot n_j = n_{jE}$ and $(1-r) \cdot n_j = n_{jC}$ with $0 \leq r \leq r$ for all $1 \leq j \leq K$ becomes

$$r(1-r)N = \frac{\sigma^2}{\Delta^2}(t_{N-2K}(1-\beta) + t_{N-2K}(1-\alpha/2))^2 \tag{9}$$

This formula, derived under the assumption of homogenous variances using the optimal weights and the allocation ratio of r, can be evaluated under various perspectives. One can determine the sample size necessary to detect a certain treatment effect of a clinical trial or to determine the power for various settings of the allocation ratio. Of course, the relationship of the sample size to the RP is obvious in the case of RPs forcing terminal balance. The power can also be related to RPs with the maximal tolerated imbalance margin a. The margin can be justified on the basis of the tolerable loss in power resulting from unbalanced allocation. In this case, equation (9) can be used to describe the relationship between r and the power. Both aspects are mentioned in the numerical evaluation study below. Using the weights $w_j = 1$, the left-hand side of equation (8) yields $\frac{r(1-r)K^2}{\sum_{j=1}^{K} \frac{1}{n_j}}$ and thus depends on the center sample sizes. In the case of equal center sample sizes, the same formula can be used for the unweighted test. In contrast to the weighted test, the sample size formula for the unweighted case requires assumptions if unbalanced sample sizes across centers are assumed.

4 Stratification in the presence of bias

We now turn to the question of bias. Two common forms of bias encountered in clinical trials are *chronological bias* due to time trends in patient outcomes,[16] and *selection bias*, which can result in covariate imbalances and inflation of type I error rates.[3] By definition, selection bias arises from the conscious or unconscious guessing of treatment assignments so that patients have a higher chance of assignment to the investigator's treatment of choice for those patients. While double-blinded studies, and multi-center studies with a central randomization unit mitigate the possibility of selection bias, Berger[17] gives numerous examples of when selection bias has arisen in practice. As the ICH E9 Guidelines note,[12]

> It is important to identify potential sources of bias as completely as possible so that attempts to limit such bias may be made.... The treatment effect and treatment comparisons should involve consideration of the potential contribution of bias to the p-value.

A recent paper provides a template on assessing the potential for chronological or selection bias and gives guidance on how to choose an appropriate RP and test statistic to account for that possibility.[6] Here, we use a similar model to determine the impact on Fleiss's test in the presence of such bias.

We first specify a compound bias vector τ_{ji} for stratum j and patient i that is a linear combination of a metric of chronological bias and selection bias. Taking into account the stratified randomization, we explore a linear time-trend[16] model per stratum similar to Hilgers[6] given by

$$\tau_{ji} = \underbrace{\theta_j \frac{i}{n_{jE}+n_{jC}}}_{\text{linear time trend}} + \underbrace{\eta_j \frac{n_{jE}(i-1)-n_{jC}(i-1)}{n_{jE}(i-1)+n_{jC}(i-1)}}_{\text{selection bias}} \tag{10}$$

Hereby, the magnitude θ_j of the linear time trend varies between centers. Note that Hilgers[6] proposed to formulate θ_j as fraction of the variance σ^2. The second term generalizes the biasing policy first introduced by Proschan[3] for the Gauss test and later investigated by Hilgers[6] for the t test. The amount of selection bias $\eta_j \geq 0$ is allowed to vary between centers. The biasing policy in equation (10) "favors" or biases the expected response towards treatment E assuming if the less frequent treatment allocated so far is E assuming E will be allocated next. The direction $\eta_j \geq 0$ corresponds to favoring E. Other metrics have been used to define the selection bias metric, including just the sign of $n_{jE}(i-1) - n_{jC}(i-1)$. We chose our metric so that it is roughly the same scale as the chronological bias metric.

5 Evaluation criterion

In our numerical evaluation study, we enumerate all possible randomization sequences for four different procedures and compare the bias to the type I error rate via computing the proportion of sequences that preserve the type I error rate at the nominal (0.05) level. If there is no bias (e.g. $\eta_j = \theta_j = 0$), 100% of sequences will preserve the type I error rate, regardless of the procedure used. To be more formal, denote the bias vector $\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{1n_j}, \ldots, \tau_{K1}, \ldots, \tau_{Kn_j})$ and the set of all sequences $\mathbf{z}$ generated by the RP by Ω_{RP}. The test statistic $t(\mathbf{z})$ depends on the randomization sequence is central t distributed with $N-2k$ degrees of freedom under the null hypotheses and no bias $\boldsymbol{\tau} = \mathbf{0}$, i.e. the null hypotheses H_0 will be rejected at the α level if $|t(\mathbf{z})| \geq t_{N-2k}(1-\alpha/2)$. Then, the evaluation criterion can be expressed by using our distributional result above including the non-centrality parameter (7)

$$\begin{aligned} P_{RP,\boldsymbol{\tau}}(H_1|H_0) &= P_{RP,\boldsymbol{\tau}}\Big(\mathbf{Z} \in \{0,1\}^N : \Big|t_{N-2k,\delta(\mathbf{Z}),\lambda(\mathbf{Z})}\Big(1-\frac{\alpha}{2}\Big)\Big| \geq \Big|t_{N-2k}\Big(1-\frac{\alpha}{2}\Big)\Big|\Big) \\ &= \sum_{\mathbf{z}\in\Omega_{RP}} 1\Big\{F_{N-2k,\delta(\mathbf{z}),\lambda(\mathbf{z})}\Big(t_{N-2k}\Big(\frac{\alpha}{2}\Big)\Big) + F_{N-2k,-\delta(\mathbf{z}),\lambda(\mathbf{z})}\Big(t_{N-2k}\Big(\frac{\alpha}{2}\Big)\Big) \leq \alpha\Big\} P_{RP,\boldsymbol{\tau}}(\mathbf{Z}=\mathbf{z}) \end{aligned} \tag{11}$$

where $F_{N-2k,\delta,\lambda}$ denotes the distribution function of the doubly non-central t-distribution with $N-2K$ degrees of freedom and non-centrality parameters $\delta(\mathbf{Z})$ and $\lambda(\mathbf{Z})$. In the ideal case, the probability should be 1, meaning that the 5% level is maintained by all allocation sequences. A value below 1 indicates that the actual type I error rate is higher than the target level of 5%. Note that this quantity summarizes the impact of bias over all randomization sequences and demonstrates the clinical consequences as well as the "go/no-go" decision of the regulator directly.

6 Numerical evaluation study

The objective of the following numerical evaluation study is to illustrate effects of stratification in both the randomization and the test statistic. It is not intended to conduct a comprehensive simulation study, recognizing that the specification of the sample size as well as θ_j and η_j depends on the practical situation. To be more specific, we start with a $K=2$ center clinical trial and use a total sample size of 80 patients with common θ_j and η_j in all centers. The following reasoning leads to the specification of θ_j and η_j. Concerning the linear time trend θ_j, it should be noted that although the θ_j are defined within each center, the maximal extent of the time trend should not exceed σ. In contrast, although the magnitude of the selection bias effect η_j may vary between centers, it is like a population effect within center and no maximal extent restriction may apply. To relate the total sample size of 80 in a $K=2$ center clinical trial to the effect size, formula (9) is used. The hypothesis $H_0 : \mu_E = \mu_C$ vs. $\tilde{H}_1 : \mu_E - \mu_C = \Delta$ should be tested with the (optimal) weighted t test (equation (3)) assuming common variance $\sigma = 1$ and intended allocation ratio of 1:1 at the 5% significance level with a power of 80%. This results in a uniform effect size of $\Delta = 0.635$. With this effect size, the allocation ratio r is varied so that the loss in power does not exceed 2%. This yields an allocation ratio of $r = 0.608$ which translates to sample size of 31:49 corresponding to a maximal tolerable imbalance of 18. With the uniform or proportional spread, this results in a maximal tolerated imbalance by center of 4 and 5, respectively.

For illustration purposes, we will compare the stratified and unstratified versions of CR, BSD(9), PBR(4), and EBC(2/3). These four procedures represent complete randomization and the three types of restricted randomization mentioned earlier. These procedures were evaluated for two different splits of the total sample size ($n_1 = n_2 = 40$ and $n_1 = 60, n_2 = 20$) and the combinations of selection and time-trend bias as $(\eta, \theta) = (0, 0.2), (0.2, 0), (0.2, 0.2), (0, 0.05), (0.05, 0), (0.05, 0.05)$. The evaluation criterion was the number of sequences protecting the 5% level for stratified and unstratified randomization as well as stratified ($w_j = 1, w_j^*$) and unstratified (*us*) test statistic and RP (see Supplementary Material). Note that unstratified randomization and test statistic correspond to the case presented in Hilgers.[6] The results for (0,0.05), (0.05,0), (0.05,0.05) are given in Table 1 as well as for (0,0.2), (0.2,0), (0.2,0.2) in Table 2. In an additional evaluation, the number of centers K is increased from 2 to 8 while splitting the total sample uniformly to the centers to show, whether there is a different influence on the type I error rate. We used an R software script to conduct the analysis, see Supplementary Material.

In the case where both biases are present, the stratified randomization with stratified analysis performs worse than unstratified analysis scenarios. The magnitude does not depend on the balancing of sample sizes between centers (20 : 60 vs. 40 : 40; Table 1). Using the favored weighted test statistic following a stratified analysis, it appears that BSD and CR perform much better than all other RPs in the both biased scenarios. However, the effect depends markedly on the type of bias. In the case of only time trend in the data, the final balance procedures (EBC(0.67), PBR) perform better than BSD or CR as well as with the unstratified analysis following unstratified randomization. Weightig with w_j^* performs uniformly better than weighting with $\mathrm{w_j} = 1$.

7 Discussion

The approach presented in this paper for multi-center trials follows the ideas of the evaluation of randomization procedures for design optimization (ERDO)[6] framework. However as outlined, many aspects need to be addressed to demonstrate the contribution of randomization in mitigating bias during the planning phase of a multi-center trial.

Although Kraemer[18] discussed various RPs in clinical trials including stratification, the most common choice of stratified randomization is PBR with common block size.[19–22] We have presented new aspects to formulate RPs, whether unrestricted or restricted, in order to induce the final balance or maximal tolerated imbalance including PBR in a stratified form. We have discussed the formulation of stratified unrestricted and restricted procedures forcing balance in probability, forcing balance by maximal tolerable imbalance, and forcing terminal balance as three subclassifications of restricted RPs.

There are several limitations of this study. First, our compound criterion for selection bias and chronological bias imposes similar scaling, but it is difficult or impossible to scale them identically. Second, the weighting of the two criteria is subjective and may be adjusted to account for the different scaling. Although our statistical test assumes homogeneous variances across centers, the methodology can be used with standardized observation in the case of known heterogeneous variances across centers.

Our proposed approach is demonstrated in a numerical evaluation study. Here, we use very specific settings, e.g. common selection bias and time-trend effects across centers, limited sample sizes corresponding to a particular effect size. We are aware that this evaluation study does not mirror all practical situations. However, specific practical situations of the multi-center clinical trial to be planned can be embedded easily into the evaluation study to demonstrate the corresponding effects. Moreover, the corresponding results for different evaluation metrics, e.g. mean type I error probability, are supplemented in tables. We used the supplemented R code for all computations.

We have chosen to use a parametric t test as our evaluation statistic rather than the more natural randomization test.[23] Randomization tests can be computed easily through the Monte Carlo re-randomization methodology, although power considerations are computationally intensive. They tend to preserve type I error rates under time trends and have no distributional assumptions.[2] Randomization tests can be formulated easily incorporating stratification, but the theoretical results we have derived herein would be impossible for exact randomization tests or Monte Carlo re-randomization tests.

Our theoretical derivation could be applied to a general class of weights w_j including, in particular, the inverse variance approach, although we focus our numerical evaluation study to the weights $w_j = 1$ or $w_j = w_j^*$, see Lin.[15] Lin stated that many statisticians as well as the U.S. Food and Drug Administration recommend the unweighted $w_j = 1$ analysis.

Table 1. Probability of stratified and unstratified randomization procedures to keep the 5% level for BSD(9), CR, EBC(0.67) and PBR(4) depending on the amount of selection $\eta = 0, 0.05$ and time-trend bias $\Theta = 0, 0.05$ for different allocation ratios and analysis using weighted (w_j^*), unweighted ($w_j = 1$) and unstratified (*us*) *t* test.

				Stratified randomization			Unstratified randomization		
Allocation ratio	Θ	η	Randomization procedure	w_j^* test	$w_j = 1$ test	*us*-test	w_j^* test	$w_j = 1$ test	*us*-test
20 : 60	0.05	0	BSD (9)	0.58	0.27	0.71	0.58	0.27	0.67
			CR	0.58	0.28	0.67	0.58	0.28	0.68
			EBC (0.67)	0.85	0.41	0.95	0.76	0.41	0.96
			PBR (4)	1.00	0.93	1.00	1.00	0.93	1.00
	0	0.05	BSD (9)	0.35	0.12	0.47	0.35	0.12	0.47
			CR	0.34	0.11	0.46	0.34	0.11	0.47
			EBC (0.67)	0.11	0.04	0.19	0.17	0.04	0.19
			PBR (4)	0.00	0.00	0.00	0.00	0.00	0.00
	0.05	0.05	BSD (9)	0.43	0.17	0.66	0.41	0.17	0.63
			CR	0.42	0.17	0.63	0.42	0.17	0.63
			EBC (0.67)	0.22	0.08	0.84	0.30	0.08	0.86
			PBR (4)	0.03	0.00	1.00	0.03	0.00	1.00
40 : 40	0.05	0	BSD (9)	0.57	0.31	0.76	0.59	0.31	0.67
			CR	0.59	0.32	0.68	0.59	0.32	0.68
			EBC (0.67)	0.82	0.52	0.97	0.74	0.52	0.96
			PBR (4)	1.00	1.00	1.00	1.00	1.00	1.00
	0	0.05	BSD (9)	0.35	0.16	0.47	0.34	0.16	0.47
			CR	0.34	0.16	0.47	0.36	0.16	0.47
			EBC (0.67)	0.11	0.04	0.20	0.15	0.04	0.20
			PBR (4)	0.00	0.00	0.00	0.00	0.00	0.00
	0.05	0.05	BSD (9)	0.42	0.23	0.69	0.43	0.23	0.62
			CR	0.43	0.23	0.63	0.42	0.23	0.63
			EBC (0.67)	0.22	0.10	0.88	0.29	0.10	0.87
			PBR (4)	0.03	0.00	1.00	0.02	0.00	1.00
8 × 10	0.05	0	BSD (2)	0.79	0.13	0.98	0.68	0.13	0.66
			CR	0.69	0.10	0.68	0.68	0.10	0.68
			EBC (0.67)	0.78	0.12	0.91	0.71	0.12	0.96
			PBR (2)	1.00	0.36	1.00	0.81	0.36	1.00
	0	0.05	BSD (2)	0.00	0.00	0.15	0.04	0.00	0.48
			CR	0.05	0.00	0.48	0.05	0.00	0.47
			EBC (0.67)	0.01	0.00	0.21	0.02	0.00	0.19
			PBR (2)	0.00	0.00	0.00	0.00	0.00	0.00
	0.05	0.05	BSD (2)	0.00	0.00	0.86	0.05	0.00	0.63
			CR	0.05	0.00	0.62	0.05	0.00	0.63
			EBC (0.67)	0.01	0.00	0.79	0.02	0.00	0.86
			PBR (2)	0.00	0.00	1.00	0.00	0.00	1.00

BSD: big stick design; EBC: Efron's biased coin design; PBR: permuted block randomization; CR: complete randomization.

Sample size considerations are presented by various authors. Whereas Ruvuna[24] and Vierron and Giraudeau[25] used the normal approximation formula, Lin's[15] approach is based on the *t* statistic. We presented a general sample size formula for the weighted *t* test with *K* centers which generalizes Lin's approach for the two center case and the weighted (w_j^*) and unweighted $w_j = 1$ evaluation. Among others, our results can be used to demonstrate the effect on the power when adding centers during progress of the trial, which seem to be common practice to increase recruitment. Furthermore, our formulas can be used for power considerations, when imbalance in sample sizes between centers is assumed.[24,25] Although it was not discussed in here, the approach can be extended to the case of random center size by using the corrected variance formulas of Ganju and Mehrotra.[26]

Although some authors mention that randomization is used to avoid bias, bias is quite likely to occur when the PBR is used, particularly when the block size is small. We present a general formal analytical approach to show how RPs are able to limit the impact of selection and chronological bias on the test decision.

Table 2. Probability of stratified and unstratified randomization procedures to keep the 5% level for BSD(9), CR, EBC(0.67) and PBR(4) depending on the amount of selection $\eta = 0, 0.2$ and time-trend bias $\Theta = 0, 0.2$ for different allocation ratios and analysis using weighted (w_j^*), unweighted ($w_j = 1$) and unstratified (*us*) *t* test.

			Randomization	Stratified randomization			Unstratified randomization		
Allocation ratio	Θ	η	procedure	w_j^* test	$w_j = 1$ test	*us*-test	w_j^* test	$w_j = 1$ test	*us*-test
20 : 60	0.2	0	BSD (9)	0.65	0.31	0.71	0.67	0.31	0.67
			CR	0.66	0.32	0.68	0.66	0.32	0.68
			EBC (0.67)	0.90	0.47	0.95	0.84	0.47	0.96
			PBR (4)	1.00	0.97	1.00	1.00	0.97	1.00
	0	0.2	BSD (9)	0.45	0.15	0.54	0.43	0.15	0.53
			CR	0.45	0.16	0.54	0.44	0.16	0.55
			EBC (0.67)	0.15	0.05	0.23	0.23	0.05	0.24
			PBR (4)	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	BSD (9)	0.48	0.18	0.62	0.48	0.18	0.61
			CR	0.48	0.18	0.62	0.48	0.18	0.60
			EBC (0.67)	0.28	0.09	0.83	0.34	0.09	0.84
			PBR (4)	0.03	0.00	1.00	0.03	0.00	1.00
40 : 40	0.2	0	BSD (9)	0.66	0.36	0.75	0.66	0.36	0.65
			CR	0.67	0.36	0.67	0.67	0.36	0.67
			EBC (0.67)	0.89	0.60	0.97	0.82	0.60	0.96
			PBR (4)	1.00	1.00	1.00	1.00	1.00	1.00
	0	0.2	BSD (9)	0.45	0.22	0.53	0.45	0.22	0.54
			CR	0.44	0.21	0.54	0.45	0.21	0.54
			EBC (0.67)	0.15	0.05	0.24	0.22	0.05	0.25
			PBR (4)	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	BSD (9)	0.48	0.25	0.67	0.48	0.25	0.61
			CR	0.48	0.25	0.62	0.47	0.25	0.61
			EBC (0.67)	0.27	0.11	0.86	0.33	0.11	0.84
			PBR (4)	0.01	0.00	1.00	0.01	0.00	1.00
8 × 10	0.2	0	BSD (2)	0.72	0.11	0.97	0.61	0.11	0.66
			CR	0.61	0.08	0.68	0.62	0.08	0.68
			EBC (0.67)	0.70	0.10	0.91	0.64	0.10	0.96
			PBR (2)	1.00	0.37	1.00	0.72	0.37	1.00
	0	0.2	BSD (2)	0.00	0.00	0.25	0.03	0.00	0.54
			CR	0.03	0.00	0.54	0.03	0.00	0.54
			EBC (0.67)	0.00	0.00	0.26	0.01	0.00	0.24
			PBR (2)	0.00	0.00	0.00	0.00	0.00	0.00
	0.2	0.2	BSD (2)	0.00	0.00	0.82	0.04	0.00	0.60
			CR	0.04	0.00	0.61	0.03	0.00	0.61
			EBC (0.67)	0.00	0.00	0.77	0.01	0.00	0.85
			PBR (2)	0.00	0.00	1.00	0.00	0.00	1.00

BSD: big stick design; EBC: Efron's biased coin design; PBR: permuted block randomization; CR: complete randomization.

The idea behind the selection bias used originates from a natural preference for one of the treatments. Furthermore, it seems to be very common, assuming that the allocation process tends to produce a balanced allocation ratio at least at the end, that investigators would believe that the treatment used most frequently thus far is less likely to appear next. Combining these two arguments, it may be reasonable, that in the situation of knowledge or best guessing what the next allocation would probably be, to choose the next patient according to the expected next treatment. This is also in line with the patient's hope to be assigned to the better treatment. Summarizing, it has to be stated that this process is unconscious or subconscious. The question is not whether selection bias occurs or not, but rather how much impact of bias one is willing to accept. This can be investigated with the proposed sensitivity analysis approach even in the planning phase. With this consideration, a unique approach is presented to link the randomization process of unrestricted or restricted procedures with the trial outcome.

Of course, other biases for time trend, e.g. log-time trend and step time trend[16] or attrition bias could be easily implemented in the modeling and then used in a numerical evaluation study. For instance, attrition bias could be modeled by a variable taking 0 or 1 on missingness, which offers opportunities, to study mechanism like missing at random.

Within this paper, we formulate a biasing policy for selection and chronological bias for a two-arm, parallel group, multi-center trial, according to the weighted stratified t test procedure proposed by Fleiss.[1] We further derive the distribution of the stratified weighted test statistic to calculate the impact on the type I error rate. Finally, the impact of the combined additive bias in multi-center trials using the unstratified t test compared to the weighted stratified t test is demonstrated in a simulation study.

8 Conclusion

Stratification in the randomization process makes the analysis sensitive to bias, i.e. results in type I error inflation. Procedures forcing terminal balance are worse in the cases where the study is prone to selection bias, irrespective if time trend is present additionally. Unbalanced sample size between centers does not affect the results. This leads to the conclusion that stratification in the randomization should be considered carefully if bias is supposed to be present. In summary, the presented approach contributes to optimizing the design of clinical trials stratified by center with respect to improve the derived level of evidence.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the IDeAl project funded from the European Union Seventh Framework Programme (FP7 2007-2013) under grant agreement No. 602552. RDH received funding from the European Joint Programme on Rare Diseases within European Union's Horizon 2020 research and innovation program under grant agreement No. 825575. Part of the work was done while RDH joined 2018 workshop on Design of Experiments: New Challenges at CIRM Luminy, France. RDH was granted by RWTH Aachen University under project rwth0334 with computing resources for simulations.

Supplemental material

Supplemental material is available for this article online.

ORCID iDs

Ralf-Dieter Hilgers https://orcid.org/0000-0002-5945-1119
Nicole Heussen https://orcid.org/0000-0002-6134-7206

References

1. Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials* 1986; **7**: 267–275.
2. Rosenberger WF and Lachin J. *Randomization in clinical trials: theory and practice*. New York, NY: Wiley, 2016.
3. Proschan M. Influence of selection bias on type I error rate under random permuted block designs. *Stat Sin* 1994; **4**: 219–231.
4. Kennes LN, Cramer E, Hilgers RD, et al. The impact of selection bias on test decisions in randomized clinical trials. *Stat Med* 2011; **30**: 2573–2581.
5. Tamm M, Cramer E, Kennes LN, et al. Influence of selection bias on the test decision – a simulation study. *Methods Inf Med* 2012; **51**: 138–143.
6. Hilgers RD, Uschner D, Rosenberger WF, et al. ERDO – a framework to select an appropriate randomization procedure for clinical trials. *BMC Med Res Methodol* 2017; **17**(1): 159.
7. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* 1971; **58**: 403–417.

8. Soares JF and Wu CFJ. Some restricted randomization rules in sequential designs. *Commun Stat Theory Methods* 1982; **12**: 2017–2034.
9. Mantel N. Random numbers and experimental design. *Ann Stat* 1969; **23**: 32–34.
10. Zelen M. The randomization and stratification of patients to clinical trials. *J Chronic Dis* 1974; **27**: 365–375.
11. Berger VW, Ivanova A and Knoll DM. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. *Stat Med* 2003; **22**(19): 3017–3028.
12. ICH E9. Statistical principles for clinical trials. https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf (accessed 24 April 2019).
13. Johnson NL and Kotz S. *Continuous univariate distributions – 2*. New York, NY: Wiley, 1970.
14. Searle SR. *Linear models*. New York, NY: Wiley, 1971.
15. Lin Z. An issue of statistical analysis in controlled multi centre studies: how shall we weight the centres? *Stat Med* 1999; **18**: 365–373.
16. Tamm M and Hilgers RD. Chronological bias in randomized clinical trials arising from different types of unobserved time trends. *Methods Inf Med* 2014; **53**: 501–510.
17. Berger VW. *Selection bias and covariate imbalances in randomized clinical trials*. Chichester: Wiley, 2005.
18. Kraemer H and Fendt KH. Random assignment in clinical trials: issues in planning (infant health and development program). *J Clin Epidemiol* 1990; **43**: 1157–1167.
19. Ganju J and Zhou K. The benefit of stratification in clinical trials revisited. *Stat Med* 2011; **30**: 2881–2889.
20. Pickering RM and Weatherall M. The analysis of continuous outcomes in multi-centre trials with small centre sizes. *Stat Med* 2007; **26**: 5445–5456.
21. Chu R, Thabane L, Ma J, et al. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC Med Res Methodol* 2011; **11**: 21.
22. Feaster DJ, Mikulich-Gilbertson S and Brinks AM. Modeling site effects in the design and analysis of multisite trials. *Am J Drug Alcohol Abuse* 1998; **37**: 383–391.
23. Zheng L and Zelen M. Multi-center clinical trials: randomization and ancillary statistics. *Ann Appl Stat* 2008; **2**(2): 582–600.
24. Ruvuna F. Unequal center sizes, sample size, and power in multicenter clinical trials. *Drug Inf J* 2004; **38**: 387–394.
25. Vierron E and Giraudeau B. Sample size calculation for multicenter randomized trial: Taking the center effect into account. *Control Clin Trials* 2007; **28**: 451–458.
26. Ganju J and Mehrotra DV. Stratified experiments reexamined with emphasis on multicenter trials. *Control Clin Trials* 2003; **24**: 167–181.