Chronological Bias in Randomized Clinical Trials Arising from Different Types of Unobserved Time Trends

M. Tamm; R.-D. Hilgers
Department of Medical Statistics, RWTH Aachen University, Aachen, Germany

Keywords
Chronological bias, time trends, drift, permuted block randomization

Summary
Background: In clinical trials patients are commonly recruited sequentially over time incurring the risk of chronological bias due to (unobserved) time trends. To minimize the risk of chronological bias, a suitable randomization procedure should be chosen.

Objectives: Considering different time trend scenarios, we aim at a detailed evaluation of the extent of chronological bias under permuted block randomization in order to provide recommendations regarding the choice of randomization at the design stage of a clinical trial and to assess the maximum extent of bias for a realized sequence in the analysis stage.

Methods: For the assessment of chronological bias we consider linear, logarithmic and stepwise trends illustrating typical changes during recruitment in clinical practice. Bias and variance of the treatment effect estimator as well as the empirical type I error rate when applying the t-test are investigated. Different sample sizes, block sizes and strengths of time trends are considered.

Results: Using large block sizes, a notable bias exists in the estimate of the treatment effect for specific sequences. This results in a heavily inflated type I error for realized worst-case sequences and an enlarged mean squared error of the treatment effect estimator. Decreasing the block size restricts these effects of time trends. Already applying permuted block randomization with two blocks instead of the random allocation rule achieves a good reduction of the mean squared error and of the inflated type I error. Averaged over all sequences, the type I error of the t-test is far below the nominal significance level due to an overestimated variance.

Conclusions: Unobserved time trends can induce a strong bias in the treatment effect estimate and in the test decision. Therefore, already in the design stage of a clinical trial a suitable randomization procedure should be chosen. According to our results, small block sizes should be preferred, but also medium block sizes are sufficient to restrict chronological bias to an acceptable extent if other contrary aspects have to be considered (e.g. serious risk of selection bias). Regardless of the block size, a blocked ANOVA should be used because the t-test is far too conservative, even for weak time trends.

1. Introduction

Clinical trials should be carefully designed and conducted to ensure the comparability of treatment groups and thus minimize bias in the study results. One kind of bias is chronological bias [1, 2] which occurs if treatment effects are confounded with time trends. It is a special form of accidental bias [3] when considering a time-heterogeneous covariate. Time trends are most likely in studies with a long recruitment phase due to high sample sizes or due to slow accrual rates. Especially in studies on rare diseases, slow accrual rates are common [4]. Time trends can arise for different reasons: changes in the covariate distribution due to changed recruiting criteria (e.g. amendment of inclusion / exclusion criteria [5 (p 10), 6], changes in staff) or increasing referrals, learning effects in the application of e.g. surgical methods [7], improvements of the concomitant therapy or of diagnostic methods [8], or changes in the disease itself. Thus, trends can be based on changes in the patient population or on changes in the therapy. Various studies have demonstrated the existence of time trends in clinical trials [cf. 9 –12].

One goal of randomization is to achieve balance between the treatment groups throughout the recruitment time of the clinical trial to minimize chronological bias. Nevertheless, chronological bias can occur if, by chance, a long series of consecutive patients receives the same treatment. If in such a situation, for example patients with a worse prognosis are included at the beginning and patients with a better prognosis at the end of the trial, the time trend is likely to impact the treatment effect estimate. To overcome these disadvantages, the ICH E9 guideline [6] explicitly recommends the usage of randomization.
in blocks [13]. The restriction on the randomization achieves a better balance over time and should therefore minimize chronological bias. However, every restriction increases the predictability of the next assignments and therefore increases vulnerability to selection bias [e.g. 14–17]. The use of small block sizes is not advisable with respect to selection bias. Hence, there exists a trade-off between selection bias and chronological bias.

To get further insight into this problem and to give more explicit recommendations, a detailed assessment of chronological bias is necessary, especially regarding the benefit of using small block sizes. Although several authors have used the maximum imbalance between groups over the course of a trial to describe the impact of the randomization procedure on chronological bias [e.g. 2, 18], a formal approach to quantify the influence of time trends on the results of a clinical trial was not given. Efron [3] introduced an approach to quantify accidental bias, though, this is only a worst case approximation and may not reflect the properties in practice, especially when considering a time-heterogeneous covariate, as noted by Rosenberger and Lachin [13, section 5.6]. They quantified the impact of time trends through the probability of a covariate imbalance, but only a specific time trend was examined in their simulations.

Since in clinical trials the treatment effect estimates and p-values are of interest, we extend these results focusing on the bias and mean squared error of the treatment effect estimator and on the empirical type I error rate. In Kalish and Begg [10], simulation results concerning the empirical type I error rate for a linear trend and a binary outcome are given. They concluded that rather than the time trend, the discrete nature of the exact test for a binary outcome variable causes a distorted nominal significance level. This is not an issue for a continuous outcome, which is considered in the following. A continuous outcome was studied by Rosenkranz [19], who described the impact of different randomization procedures on the analysis of clinical trials when an unobserved covariate is present. By way of illustration, Rosenkranz simulated a linear time trend to show the impact of the randomization on the results of the study. In our investigation, these results have partly been used and extended whereas we consider another randomization procedure, the widely used permuted block randomization, and various kinds of time trends. Because the objective of this paper is to assess and quantify chronological bias, we focus not only on an overall evaluation (to give recommendations regarding the selection of the most suitable randomization in the planning stage) but also include worst-case evaluations for realized allocation sequences (to address the possible bias for a given sequence in the analysis stage). Using different formulations of the time trend as an unknown covariate, we achieve concrete results, upper limits and explicit recommendations regarding chronological bias.

In section 2, we describe the model and the types of time trends considered herein. Section 3 has two parts: In subsection 3.1, bias and variance of the treatment effect estimator are derived and quantified for different block sizes and individual sequences. In subsection 3.2, the impact on statistical test decisions is given, using simulations concerning type I error. These results are supported by theoretical considerations regarding the variance estimation.

2. Methods
2.1 Model

We consider a clinical trial where patients are sequentially enrolled in a two arm parallel group design using permuted block randomization. The outcome is measured on a continuous scale. We assume that the number of patients $n$ is even and $n/2$ patients are allocated to a new treatment A and $n/2$ patients are allocated to a placebo/active comparator B. The patients are consecutively numbered from 1 to $n$ according to their time of inclusion in the trial. According to Rosenkranz [19], the response $Y_i$, $i = 1, \ldots, n$, can be described by the linear model

$$Y_i = \mu_A + Z_i + \mu_B \cdot (1 - Z_i) + \tau(i) + \varepsilon_i,$$

where $\mu_A$ and $\mu_B \in \mathbb{R}$ are the expected effects under treatment A and B, respectively. The independent random errors $\varepsilon_i$, $i = 1, \ldots, n$, are normally distributed with mean 0 and unknown variance $\sigma^2$. The random vector $Z = (Z_1, \ldots, Z_n)$ represents the randomization sequence with $Z_i = 1$ if the $i$-th patient is allocated to treatment A and $Z_i = 0$ if the $i$-th patient is allocated to treatment B and realizations $z = (z_1, \ldots, z_n) \in \{0,1\}^n$. The random errors $\varepsilon_i$ are assumed to be independent of $Z = (Z_1, \ldots, Z_n)$. In our considerations, the covariate $t(i) \in \mathbb{R}$ describes the time trend due to the sequential arrival of the patients. To cover a wide range of practical situations, as discussed in the introduction, we allow for linear, stepwise and logarithmic time trends and specify the covariate as follows:

i. Linear trend: $\tau(i) = \lambda \cdot (i - 1)$ with $\lambda \in \mathbb{R}$

ii. Stepwise trend: $\tau(i) = \lambda \cdot \lfloor i \rfloor \geq c$ (2.1) with $\lambda \in \mathbb{R}$ and $c \in \mathbb{N}$, $1 \leq c \leq n$

iii. Logarithmic trend: $\tau(i) = \lambda \cdot \log(i)$ with $\lambda \in \mathbb{R}$

In this setting, $\lambda \in \mathbb{R}$ describes the strength of the trend increasing with the consecutive patient number. A linear trend could e.g. reflect steady improvement in the concomitant therapy, which results in a shifted response by $\lambda$ with every patient included. An amendment due to, e.g., a necessary change of inclusion criteria could be reflected by a stepwise trend with a change by $\lambda$ after the amendment. A logarithmic trend could describe learning effects, which diminish with the number of patient treated. If the patients are allocated uniformly over time, the consecutive patient number may also be interpreted in terms of recruitment time $t_i \in \mathbb{R}_{\geq 0}$. Moreover, e.g. model (iii) approximates other recruitment patterns such as slow catch up [20] described by recruitment time $t_i = \log(i)$ in combination with a linear trend $\tau^*(t) = \lambda \cdot t_i$ over recruitment time $t_i$.

2.2 Permuted Block Randomization and Random Allocation Rule

Throughout this paper, we focus on the permuted block randomization (PBR) de-
scribed in Rosenberger and Lachin [13] with fixed and even block sizes k and number of blocks M, i.e. n = M · k. This also includes the random allocation rule (RAR) with one (M = 1) block of size k = n. The set of all possible sequences for the permuted block randomization is denoted by Γ(PBRnk) [2]. Transferring the properties of the RAR given in [13 (chapter 3.3)] to the PBR, it is E(Zi) = 1/2 and Var(Zi) = 1/4 for the allocations Zi, i = 1, ..., n, of the random sequence Z = (Z1, ..., Zn) with realization z = (z1, ..., zn) ∈ Γ(PBRnk). For the covariance structure within one block we have Cov(Zi, Zj) = -1/(4(k – 1)) and E(ZiZj) = (k – 2)/(4(k – 1)), whereas between blocks we have Cov(Zi, Zj) = 0 and E(ZiZj) = 1/4. Since the allocations are balanced within one block, it is

\[
\sum_{i \in B_{k}} Z_i - \sum_{i \in B_{k}} (1 - Z_i) = 0
\]

with Bi = [i ∈ N | 1 + (r – 1) · k ≤ i ≤ r · k], r = 1, ..., M. These notations and properties are used in section 3.

2.3 Assessment of Chronological Bias

To assess the impact of time trends on the results of a clinical trial, we consider the bias and the variance of the treatment effect estimator. Furthermore, the pooled variance estimate, which is used in the test statistic of the t-test, is considered. Estimates for the empirical type I error rate of the t-test under the different time trends were obtained via simulations. The simulations were conducted and analyzed using SAS 9.2 (SAS Institute Inc., Cary, NC, USA) under Windows 7. The RANUNI and RANNOR statements were used to generate pseudo random numbers for the generation of randomization sequences and the patient data, respectively. Independent starting seeds were chosen for the different scenarios. The empirical type I error rate was calculated by simulating one million data sets for each scenario under the null hypothesis of no treatment effect (\(μ_A = μ_B = 0\)) and computing the proportion of p-values that are less than the nominal 5% significance level [21] when applying the two-tailed t-test with the pooled variance estimation. For comparison, we also calculated the empirical type I error rate using blocked analysis of variance (ANOVA) which takes the block structure into account.

To get an overall evaluation of the randomization procedure, we consider the mentioned properties over all possible sequences. Furthermore, worst case scenarios are considered.

### 2.4 Choice of Parameters

In the presented numerical examples and simulations, the strength λ of the time trend is related to the residual variance in order to get a plausible magnitude of the time trend and to get comparable results between the different forms of trends and the different sample sizes. For a residual variance of \(σ^2 = 1\), the extent of the time trend is chosen to be within the one-fold and two-fold (\(v = 1\) or \(v = 2\)) standard deviation. That is, \(λ = v/(n – 1)\) for a linear trend, \(λ = v\) for the step and \(λ = v/log(n)\) for a logarithmic trend. The natural logarithm is used here. The chosen values of λ describe a change of the expected patient response by one (for \(v = 1\)) or two (for \(v = 2\)) from the first to the last patient. A total sample size of \(n = 128\) was chosen, which is reasonable in practice because using this sample size the two-tailed t-test with a 5% significance level will have 80% power to detect a difference if a medium effect size of 0.5 is present (if no time trend exists). Because time trends are quite likely in trials studying rare diseases [4], we also consider a sample size of \(n = 16\). In this case the two-tailed t-test with a 5% significance level will have 80% power to detect a difference if an effect size of about 1.5 is present. For illustration purposes, we also evaluated the case of \(n = 4\), which results in only six different sequences for the random allocation rule.

Figure 1 shows the different trends for a simulated study of \(n = 128\) with no treatment effect and \(v = 1\). Plotting the patient response against the patient index, the time trend and particularly its shape are hard to identify. The trend becomes more obvious when using the CUSUM plot suggested by Altman and Royston [9]. Even so, the correct shape remains hard to identify.

### 3. Results

#### 3.1 Estimation of the Treatment Effect

We first consider the impact of the time trend on the treatment effect estimator for the permuted block randomization. Let

\[
\bar{Y}_A = \frac{1}{n/2} \sum_{i=1}^{n/2} Z_i \cdot Y_i
\]

and

\[
\bar{Y}_B = \frac{1}{n/2} \sum_{i=1}^{n/2} (1 - Z_i) \cdot Y_i
\]

be the mean treatment effects in group A and group B, respectively. Then, the estimator of the treatment difference \(\Delta = \mu_A - \mu_B\) is given by

\[
\Delta = \bar{Y}_A - \bar{Y}_B = \frac{1}{n/2} \sum_{i=1}^{n/2} (2Z_i - 1) \cdot Y_i
\]

Given a randomization sequence \(Z = (Z_1, ..., Z_n)\), the conditional expected value for the treatment effect estimator is

\[
E(\Delta | Z) = \Delta + b(Z)
\]

Here \(b(Z)\) describes the bias in the estimate of the true treatment difference for a given sequence \(Z\). To identify the sequence which yields the maximum amount of bias, we can use the fact that all considered time trends are monotone. Obviously, for a monotonic trend the absolute amount of bias becomes maximal if for each block all allocations to treatment group A are in the first or all allocations to treatment group B, respectively. In the second half of the block or vice versa, i.e. \((z_{(r – 1)k + 1}, ..., z_{(r – 1)k + k/2}, z_{(r – 1)k + k/2 + 1}, ..., z_{2k}) = (1, ..., 1, 0, ..., 0)\) for all \(r \in \{1, ..., M\}\) or \((z_{(r – 1)k + 1}, ..., z_{(r – 1)k + k/2}, z_{(r – 1)k + k/2 + 1}, ..., z_{2k}) = (0, ..., 0, 1, ..., 1)\) for all \(r \in \{1, ..., M\}\).

In the following, we will refer to these sequences as worst-case sequences.
For the linear trend in (2.1), the worst-case sequences induce the same amount of bias in each block. Thus, we get a maximal absolute amount of

\[
\max_{x \sim F_{\text{Poisson}}(\lambda)} |b(z)| = \frac{\lambda \cdot k}{2}.
\]

Hence, we have a proportional relationship between the block size and the bias, e.g. halving the block sizes leads to a halving of the maximum extent of bias.

For the step trend \(\tau(i) = \lambda \cdot 1 \text{ if } i \geq c\), the bias only depends on the block where the step occurs and on the position of the step in this block. The allocation sequences of all other blocks do not cause a bias. Thus, with a step at \(c = (r_0 - 1) \cdot k + c_0\), i.e. at position \(c_0 \in \{1, \ldots, k\}\) in block \(r_0 \in \{1, \ldots, M\}\), the bias in ◀ Equation (3.1) is

\[
b(Z) = \frac{\lambda}{n/2} \sum_{i=1}^{m} (2Z_{c_0-1,k,i} - 1).
\]

With this we obtain an absolute maximum amount of bias for the worst-case sequences of

\[
\max_{x \sim F_{\text{Poisson}}(\lambda)} |b(z)| = \begin{cases} \frac{\lambda}{n/2} (k - (c_0 - 1)) & \text{if } c_0 > k/2, \\ \frac{\lambda}{n/2} (c_0 - 1) & \text{if } c_0 \leq k/2, \end{cases}
\]

which attains a maximum of \(\lambda \cdot k/n\) for a step at \(c_0 = k/2 + 1\). If we have small block sizes in combination with a large sample size, this time trend has no high impact. Obviously, if the allocations are balanced after the step within the block, the bias is zero.

Considering the logarithmic trend introduced in (2.1), for the worst-case sequences we have

\[
\max_{x \sim F_{\text{Poisson}}(\lambda)} |b(z)| = \frac{\lambda}{n/2} \sum_{i=1}^{m} \log \left( \frac{(rk)!(k-1)!}{(rk-1)!} \right).
\]

For the random allocation rule with \(k = n\) and \(M = 1\), the formula simplifies to

\[
\max_{x \sim F_{\text{Poisson}}(\lambda)} |b(z)| = \frac{\lambda}{n/2} \log \left( \frac{n}{n/2} \right).
\]

In ◀ Table 1, the maximum amount of bias \(b(z)\) is numerically demonstrated for different block sizes and different sample sizes. As expected, the maximum amount of bias is reduced by decreasing the block size. For a linear trend, we get half the amount of bias when using half the block size. Comparing the linear trend and the logarithmic trend, it can be observed that for large block sizes the amount of bias is larger for the linear trend than for the logarithmic trend, whereas it is almost equal or reversed for smaller block sizes. This is due to the flattening of the curve for a logarithmic trend, which has a positive impact for large block sizes. For a step function the bias greatly depends on the position of the step in the block. It is demonstrated in ◀ Table 1 for a step after half and after three quarters of the recruited patients.

From the assessment of the bias for the worst-case sequences we now proceed to the evaluation of the randomization procedure as a whole. For this, the unconditional expected value over all possible sequences for the treatment effect estimator is of interest. Rosenkranz [19] showed that for all randomization procedures with balanced groups and \(E(Z) = 1/2\), the estimate of the treatment effect is unbiased, i.e. \(E(\Delta) = \Delta\) even if an arbitrary (unknown) covariate is present. In ◀ Table 2, we illustrate all possible sequences for \(n = 4\). Since reversed sequences (e.g. 0011 vs. 1100) result in biases that are exactly the negative of each other, we obtain a total bias of zero in expectation. Instead of the expected bias we therefore use the mean squared error or the variance of the treatment effect estimator to describe the precision of the treatment effect estimator. According to Rosenkranz [19], the variance is given by the formula depicted in ◀ Figure 2 for randomization procedures with balanced groups of \(n/2\) and \(E(Z) = 1/2\). Using the

![Figure 1](https://www.methods-online.com) Patient response and accumulated patient response (CUSUM) versus consecutive number of patients (\(n = 128\, \mu_t = \mu_b = 0\, \sigma = 1\)) for different types of time trends (no time trend, linear trend with \(\lambda = 1/(n-1)\), step with \(\lambda = 1\) at \(c = 65\), logarithmic trend with \(\lambda = 1/\log(n)\))
Table 1  Results for the worst-case sequence with allocations (1, ..., 1, 0, ..., 0) in each block: Bias b(z) of treatment effect estimate as well as expected value of estimated squared standard error E(\(\hat{\gamma}_p\)\(\mid z\)) and empirical type I error rate using t-test and blocked ANOVA at nominal significance level of 0.05 (\(\mu_A = \mu_B = 0, \sigma = 1\))

<table>
<thead>
<tr>
<th>Strength of trend</th>
<th>Sample size n</th>
<th>Block size k</th>
<th>b(z) (\mid z)</th>
<th>E((\hat{\gamma}_p)(\mid z))</th>
<th>Type I error rate (t-test)</th>
<th>Type I error rate (ANOVA)</th>
<th>b(z)</th>
<th>E((\hat{\gamma}_p)(\mid z))</th>
<th>Type I error rate (t-test)</th>
<th>Type I error rate (ANOVA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear ((\lambda = u/(n - 1)))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(u = 1)</td>
<td>4</td>
<td>4</td>
<td>-0.67</td>
<td>1.0556</td>
<td>0.0670</td>
<td>0.0670</td>
<td>-0.65</td>
<td>1.0733</td>
<td>0.0643</td>
<td>0.0643</td>
</tr>
<tr>
<td>2</td>
<td>1.2222</td>
<td>0.0449</td>
<td>0.0527</td>
<td>-0.35</td>
<td>1.2195</td>
<td>0.0453</td>
<td>0.0522</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>0.53</td>
<td>0.2567</td>
<td>0.1624</td>
<td>0.1624</td>
<td>-0.43</td>
<td>0.2587</td>
<td>0.1189</td>
<td>0.1189</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.27</td>
<td>0.2719</td>
<td>0.0676</td>
<td>0.0775</td>
<td>-0.25</td>
<td>0.2672</td>
<td>0.0666</td>
<td>0.0724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.13</td>
<td>0.2757</td>
<td>0.0468</td>
<td>0.0567</td>
<td>-0.14</td>
<td>0.2703</td>
<td>0.0496</td>
<td>0.0569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>-0.50</td>
<td>0.0319</td>
<td>0.8010</td>
<td>0.8010</td>
<td>-0.28</td>
<td>0.0318</td>
<td>0.3360</td>
<td>0.3360</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>0.25</td>
<td>0.0334</td>
<td>0.2710</td>
<td>0.2919</td>
<td>-0.17</td>
<td>0.0322</td>
<td>0.1528</td>
<td>0.1573</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-0.06</td>
<td>0.0339</td>
<td>0.0539</td>
<td>0.0641</td>
<td>-0.06</td>
<td>0.0324</td>
<td>0.0570</td>
<td>0.0610</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.02</td>
<td>0.0339</td>
<td>0.0423</td>
<td>0.0513</td>
<td>-0.02</td>
<td>0.0324</td>
<td>0.0469</td>
<td>0.0510</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(u = 2)</td>
<td>4</td>
<td>4</td>
<td>-1.33</td>
<td>1.2222</td>
<td>0.1066</td>
<td>0.1066</td>
<td>-1.29</td>
<td>1.2931</td>
<td>0.0965</td>
<td>0.0965</td>
</tr>
<tr>
<td>2</td>
<td>-0.67</td>
<td>1.8889</td>
<td>0.0316</td>
<td>0.0608</td>
<td>-0.71</td>
<td>1.8780</td>
<td>0.0332</td>
<td>0.0595</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>-1.07</td>
<td>0.2767</td>
<td>0.4699</td>
<td>0.4699</td>
<td>-0.85</td>
<td>0.2848</td>
<td>0.3097</td>
<td>0.3097</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.53</td>
<td>0.3376</td>
<td>0.1046</td>
<td>0.1620</td>
<td>-0.50</td>
<td>0.3189</td>
<td>0.1041</td>
<td>0.1378</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.27</td>
<td>0.3529</td>
<td>0.0387</td>
<td>0.0766</td>
<td>-0.28</td>
<td>0.3313</td>
<td>0.0466</td>
<td>0.0759</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>-1.01</td>
<td>0.0339</td>
<td>0.9998</td>
<td>0.9998</td>
<td>-0.55</td>
<td>0.0335</td>
<td>0.8605</td>
<td>0.8605</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>-0.50</td>
<td>0.0400</td>
<td>0.7286</td>
<td>0.8012</td>
<td>-0.34</td>
<td>0.0350</td>
<td>0.4345</td>
<td>0.4633</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-0.13</td>
<td>0.0419</td>
<td>0.0608</td>
<td>0.1084</td>
<td>-0.11</td>
<td>0.0359</td>
<td>0.0750</td>
<td>0.0957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.03</td>
<td>0.0420</td>
<td>0.0255</td>
<td>0.0536</td>
<td>-0.03</td>
<td>0.0360</td>
<td>0.0390</td>
<td>0.0537</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step at (c = 0.5 \times n + 1) ((\lambda = u))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(u = 1)</td>
<td>4</td>
<td>4</td>
<td>-1.00</td>
<td>1.0000</td>
<td>0.0953</td>
<td>0.0953</td>
<td>-0.50</td>
<td>1.2500</td>
<td>0.0489</td>
<td>0.0489</td>
</tr>
<tr>
<td>2</td>
<td>1.5000</td>
<td>0.0311</td>
<td>0.0502</td>
<td>-0.50</td>
<td>1.2500</td>
<td>0.0493</td>
<td>0.0502</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>-1.00</td>
<td>0.2500</td>
<td>0.4626</td>
<td>0.4626</td>
<td>-0.50</td>
<td>0.2857</td>
<td>0.1255</td>
<td>0.1255</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0.3214</td>
<td>0.0285</td>
<td>0.0502</td>
<td>-0.50</td>
<td>0.2857</td>
<td>0.1251</td>
<td>0.1362</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.3214</td>
<td>0.0286</td>
<td>0.0500</td>
<td>0.00</td>
<td>0.3036</td>
<td>0.0325</td>
<td>0.0500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>-1.00</td>
<td>0.0313</td>
<td>0.9999</td>
<td>0.9999</td>
<td>-0.50</td>
<td>0.0352</td>
<td>0.7663</td>
<td>0.7663</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>0.00</td>
<td>0.0392</td>
<td>0.0286</td>
<td>0.0502</td>
<td>-0.50</td>
<td>0.0352</td>
<td>0.7663</td>
<td>0.7840</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.00</td>
<td>0.0392</td>
<td>0.0284</td>
<td>0.0501</td>
<td>0.00</td>
<td>0.0372</td>
<td>0.0328</td>
<td>0.0502</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.0392</td>
<td>0.0284</td>
<td>0.0502</td>
<td>0.00</td>
<td>0.0372</td>
<td>0.0329</td>
<td>0.0503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(u = 2)</td>
<td>4</td>
<td>4</td>
<td>-2.00</td>
<td>1.0000</td>
<td>0.2186</td>
<td>0.2186</td>
<td>-1.00</td>
<td>2.0000</td>
<td>0.0389</td>
<td>0.0389</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>3.0000</td>
<td>0.0079</td>
<td>0.0503</td>
<td>-1.00</td>
<td>2.0000</td>
<td>0.0393</td>
<td>0.0445</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>-2.00</td>
<td>0.2500</td>
<td>0.9605</td>
<td>0.9605</td>
<td>-1.00</td>
<td>0.3929</td>
<td>0.2795</td>
<td>0.2795</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0.5357</td>
<td>0.0056</td>
<td>0.0501</td>
<td>-1.00</td>
<td>0.3929</td>
<td>0.2791</td>
<td>0.3500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.5357</td>
<td>0.0054</td>
<td>0.0502</td>
<td>0.00</td>
<td>0.4643</td>
<td>0.0995</td>
<td>0.0501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>-2.00</td>
<td>0.0313</td>
<td>1.0000</td>
<td>1.0000</td>
<td>-1.00</td>
<td>0.0471</td>
<td>0.9993</td>
<td>0.9993</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>0.00</td>
<td>0.0630</td>
<td>0.0055</td>
<td>0.0496</td>
<td>-1.00</td>
<td>0.0471</td>
<td>0.9993</td>
<td>0.9997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.00</td>
<td>0.0630</td>
<td>0.0054</td>
<td>0.0499</td>
<td>0.00</td>
<td>0.0551</td>
<td>0.0095</td>
<td>0.0499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.0630</td>
<td>0.0055</td>
<td>0.0499</td>
<td>0.00</td>
<td>0.0551</td>
<td>0.0095</td>
<td>0.0499</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
properties described at the beginning of the section, for PBR we get with \( \tau = (1/k) \cdot \sum_{i \in B_r} \tau(i) \) Formula 3.2 in ▶ Figure 3.

According to the law of total variance, \( \sigma^2/n \) is the common variation due to the random error, i.e. \( \text{E}(\text{Var}(\hat{\Delta}|Z)) = \text{Var}(\text{E}(\hat{\Delta}|Z)) \). The second part of Formula 3.2 (▶ Figure 3) is the additional variation between the different sequences due to the time trend, i.e. \( \text{Var} (\text{E}(\hat{\Delta}|Z)) \). For a linear trend, we obtain

\[
\text{Var} (\hat{\Delta}) = \frac{4 \sigma^2}{n} + \frac{4 k}{n^2 (k-1)} \left( \sum_{r=1}^{M} \left( \sum_{i \in B_r} (\tau(i) - \bar{\tau})^2 \right) \right) \tag{3.2}
\]

Thus, using smaller block sizes \( k \) for a fixed sample size, the variance of the treatment effect estimator is reduced. Furthermore, it can be observed that the additional variation in ▶ Formula 3.3 due to the time trend reduces to approximately one fourth if the block size is halved. For a step trend during the recruitment at \( c = (r_0 - 1) \cdot k + c_0 \), we get

\[
\text{Var} (\hat{\Delta}) = \frac{4 \sigma^2}{n} + \frac{4 \lambda^2 k (k+1)}{n}. \tag{3.3}
\]

The variance depends on the block size \( k \) in combination with the position of the step within the block. For a given block size, the variance attains its maximum of \( 4 \sigma^2/(n + \lambda^2 k^2/(n^2 (k-1))) \) for a step at \( c_0 = k/2 + 1 \). As with the linear trend, the maximal amount of the variance decreases for smaller block sizes. If the step is at the beginning of the block, i.e. \( c_0 = 1 \), the variance in ▶ Formula 3.4 reduces to the unbiased variance \( 4 \sigma^2/n \) due to the random error, because the treatment effect is estimated correctly. For the logarithmic trend, the variance cannot be simplified to a short expression; however, explicitly calculated values for specific parameters are given in ▶ Table 3. It shows that the variance decreases with decreasing block size for the logarithmic trend as well. Comparing the different shapes of trends for large block and sample sizes, we see that the variance for the logarithmic trend is smaller than for the linear trend. For small block sizes the variance gets close to the unbiased variance \( 4 \sigma^2/n \) for all the considered shapes of trend (for \( n = 4: 4 \sigma^2/n = 1 \), for \( n = 16: 4 \sigma^2/n = 0.25 \), for \( n = 128: 4 \sigma^2/n = 0.0313 \)). Notably, even changing from random allocation rule to permuted block randomization with two blocks already achieves a good reduction of the additional variance.

### 3.2 Statistical Hypothesis Testing

In the following, the impact of the time trend on the test decision is shown using the independent two-tailed t-test under homoscedasticity. Since the residual variance \( \sigma^2 \) and the time trend are unknown, the variance of the treatment effect estimator has to be estimated. Comparing treatments with the t-test, this is usually done by pooling

\[
\hat{s}_a^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}_a)^2 \quad \text{and} \quad \hat{s}_b^2 = \sum_{i=1}^{n} (1-Z_i)(Y_i - \bar{Y}_b)^2,
\]

obtaining an estimated squared standard error of

\[
\hat{V}_0 = \frac{n}{n-2} \left( \frac{\hat{s}_a^2 + \hat{s}_b^2}{n} \right).
\]

Using the properties from the beginning of the section, under the null hypothesis \( \mu_a = \mu_b \) we obtain the expected value for the estimator of the variance of \( \hat{\Delta} \) (▶ Figure 4).

Thus, for a linear trend, we get a difference between the estimate and its true value of

\[
\text{E}(\hat{V}_0) - \text{Var}(\hat{\Delta}) = \frac{\lambda^2}{3} \frac{(n(n+1) - k(k+1))(n-1)}{n(n-2)}, \tag{3.5}
\]

i.e. the variance is always overestimated using the t-test. The difference in ▶ Formula 3.5 becomes larger if the block size decreases. ▶ Table 3 shows that a decrease in the block sizes achieves a reduction in the true variance of \( \hat{\Delta} \). However, at the same time, the overestimation of the variance using the t-test becomes worse. The reason is that the blocking is not incorporated in the variance estimation. For the random allocation rule the variance is estimated correctly.

For the step at position \( c_0 \in \{1, \ldots, k\} \) in block \( r_0 \in \{1, \ldots, M\} \) with \( c = (r_0 - 1) \cdot k + c_0 \), see Formula 3.6 (▶ Figure 5).
As above, this difference is always non-negative. It is equal to zero if the random allocation rule is used or if the step occurs with the first, second or last patient. In these cases, the variance is estimated correctly. In all other cases, it is overestimated subject to the block and the position where the step appears. However, for a step at the beginning of a block, i.e. \( c_0 = 1 \), the difference in Formula 3.6 (Figure 5) simplifies to

\[
E(\hat{\lambda}) - \text{Var}(\hat{\lambda}) = \frac{4\lambda^2}{n^2(n-2)}(n-r_o)(k(r_o - 2) + 2(c_0 - 1)) + \frac{(n-k)(k-c_0 + 1)(k-c_0)}{(k-1)}
\]

which is maximal for \( c = n/2 + 1 \). Obviously, in this case the difference does not depend on the block size. The properties for a logarithmic trend are demonstrated in Table 3. It behaves similarly to the linear trend, but the variance is not as heavily overestimated.

The results of the simulations concerning the type I error rate when using the t-test are given in the column “Type I error rate (t-test)” of Table 3. For the random allocation rule, the empirical type I error rate is close to the nominal 5% significance level. This coincides with the results given in Rosenkranz [19]. For small sample sizes, slight deviations in both directions are observed which might be due to the small number of possible sequences. However, for smaller block sizes and a strong linear trend, the type I error rate is far below the nominal significance level. E.g., for \( n = 128 \) and \( v = 2 \), it goes down to 0.0234 for a block size of \( k = 4 \). For a logarithmic trend, the impact is slightly weaker, which corresponds to our observation regarding the overestimation of the variance. The empirical type I error rate is 0.0359. The corresponding value for a step after one half of the recruitment is 0.0056. We find almost the same value for \( k = 64 \) and \( k = 16 \), because in all of these scenarios the step is at the beginning of a block. The situation changes for a step after three quarters of the sample size. Here, for \( k = 64 \) the step is after half of the second block and we obtain a type I error rate of 0.0343. Summarized we see that averaged over all possible sequences the test decision becomes very conservative when time trends are present and the t-test is used which does not take into account the block structure. Using the blocked ANOVA, for all scenarios we get an empirical type I error rate close to the 5% level in our simulations (Table 3).

This relation between blocked and unblocked analysis is well known from Matts and Lachin [22], who showed that for a positive intrablock correlation the analysis of variances ignoring the blocks is more conservative than the blocked analysis of variance.

While maintaining the level on average, the type I error can still be highly inflated for specific sequences, especially for the worst-case sequences. In the last column of Table 2, the empirical type I error rate when using the t-test is illustrated for \( n = 4 \). Some sequences have an inflated type I error rate and some sequences have a type I error rate below the nominal significance level, resulting in a type I error of almost five percent on average. For a given sequence, we obtain the expected value of the estimated squared standard error (Formula 3.7 in Figure 6) with

\[
\bar{\tau}_x = \frac{2}{n} \sum_{i=1}^{n} \tau(i)
\]

and \( \bar{\tau}_y = 2(2/n) \cdot \sum_{i=1}^{n} (1 - Z) \cdot \tau(i) \). Thus, the variance is overestimated for all sequences because \( \text{Var}(\hat{\lambda}|Z) = 4\sigma^2/n \). That means that using the t-test instead of the blocked ANOVA, the type I error for realized sequences is decreased due to the overestimated variance (cf. also Table 1). However, the respective worst-case sequences are penalized with the smallest overestimation, because in (3.7) the largest amount

**Table 2** Separated by individual sequences \( z \) for \( n = 4 \): Bias \( b(z) \) of treatment effect estimate as well as expected value of estimated squared standard error \( E(\hat{\lambda}|z) \) and empirical type I error rate using t-test at nominal significance level of 0.05 (\( \mu_1 = \mu_2 = 0, \sigma = 1, u = 2 \))

| \( z \) | \( b(z) \) | \( E(\hat{\lambda}|z) \) | Type I error rate (t-test) | \( b(z) \) | \( E(\hat{\lambda}|z) \) | Type I error rate (t-test) |
|---|---|---|---|---|---|---|
| Linear (\( \lambda = 2/(n-1) \)) | Logarithmic (\( \lambda = 2/\log(n) \)) |
| 1100 | -2.00 | 1.00 | 0.2189 | -1.00 | 2.00 | 0.0394 |
| 1101 | 0.00 | 3.00 | 0.0079 | -1.00 | 2.00 | 0.0396 |
| 1011 | 0.00 | 3.00 | 0.0079 | 1.00 | 2.00 | 0.0395 |
| 0111 | 0.00 | 3.00 | 0.0079 | -1.00 | 2.00 | 0.0394 |
| 0111 | 2.00 | 1.00 | 0.2183 | 1.00 | 2.00 | 0.0392 |
Table 3  Variance of treatment effect estimator \( \text{Var}(\hat{\Delta}) \) as well as expected value of estimated squared standard error \( \text{E}(\hat{\nu}) \) and empirical type I error rate using t-test and blocked ANOVA at nominal significance level of 0.05 (\( \mu_0 = \mu_1 = 0, \sigma = 1 \)) for different trends, sample sizes and block sizes

<table>
<thead>
<tr>
<th>Strength of trend</th>
<th>Sample size</th>
<th>Block size</th>
<th>( \text{Var}(\hat{\Delta}) )</th>
<th>( \text{E}(\hat{\nu}) )</th>
<th>Type I error rate (t-test)</th>
<th>Type I error rate (ANOVA)</th>
<th>( \text{Var}(\hat{\mu}) )</th>
<th>( \text{E}(\hat{\nu}) )</th>
<th>Type I error rate (t-test)</th>
<th>Type I error rate (ANOVA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Linear (( \lambda = u/(n-1) ))</td>
<td>Logarithmic (( \lambda = u/\log(n) ))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( u = 1 )</td>
<td>( u = 2 )</td>
<td>( u = 1 )</td>
<td>( u = 2 )</td>
<td>( u = 1 )</td>
<td>( u = 2 )</td>
<td>( u = 1 )</td>
<td>( u = 2 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>1.1852</td>
<td>1.1852</td>
<td>0.0502</td>
<td>0.0502</td>
<td>1.1881</td>
<td>1.1881</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>16</td>
<td>0.2752</td>
<td>0.2752</td>
<td>0.0501</td>
<td>0.0501</td>
<td>0.2703</td>
<td>0.2703</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>0.2567</td>
<td>0.2765</td>
<td>0.0426</td>
<td>0.0497</td>
<td>0.2587</td>
<td>0.2711</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>0.2519</td>
<td>0.2769</td>
<td>0.0408</td>
<td>0.0499</td>
<td>0.2535</td>
<td>0.2715</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>128</td>
<td>128</td>
<td>0.0339</td>
<td>0.0339</td>
<td>0.0502</td>
<td>0.0502</td>
<td>0.0324</td>
<td>0.0324</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64</td>
<td>0.0319</td>
<td>0.0339</td>
<td>0.0437</td>
<td>0.0503</td>
<td>0.0318</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>0.0313</td>
<td>0.0339</td>
<td>0.0415</td>
<td>0.0501</td>
<td>0.0314</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0.0313</td>
<td>0.0339</td>
<td>0.0413</td>
<td>0.0501</td>
<td>0.0313</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>128</td>
<td>0.0419</td>
<td>0.0419</td>
<td>0.0497</td>
<td>0.0497</td>
<td>0.0359</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64</td>
<td>0.0339</td>
<td>0.0420</td>
<td>0.0294</td>
<td>0.0500</td>
<td>0.0335</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>0.0314</td>
<td>0.0420</td>
<td>0.0238</td>
<td>0.0504</td>
<td>0.0317</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0.0313</td>
<td>0.0420</td>
<td>0.0234</td>
<td>0.0501</td>
<td>0.0313</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Step at ( c = 0.5 \ n + 1 ) (( \lambda = u ))</td>
<td>Step at ( c = 0.75 \ n + 1 ) (( \lambda = u ))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Discussion

The purpose of this paper was to evaluate the extent of chronological bias under different time trends models to allow for better assessment of this issue and to give recommendations for the randomization. We focused on the permuted block randomization and the two-tailed t-test, because these methods are frequently used in practice and therefore their behavior is of wide interest.

We showed that time trends can result in a strong bias in the treatment effect estimate if sequences are used which are not sufficiently balanced over time. The variance (or equivalently the mean squared error) of the treatment effect estimator can be decreased by reducing the block size. For a linear trend, e.g., the deviation from the usual variance $4\sigma^2/n$ caused by the time trend can be reduced to nearly one fourth if the block size is halved. Even when looking at the empirical type I error rate for worst-case sequences, a good effect could already be observed by a transition from the random allocation rule to the PBR with two blocks. Certainly, very small block sizes are even better with respect to chronological bias. It should be noted that for small sample sizes as in clinical trials studying rare diseases, sequences close to the worst-case sequence are not unlikely.

Based on these results, we support the statement of the ICH E9 guideline that the use of blocks is recommendable regarding chronological bias. However, small blocks increase the risk of selection bias [e.g. 14–17]. Thus, with respect to selection bias one should strike a balance, especially if the randomization list cannot be safely blinded. Since we found that already medium block sizes achieve a noticeable benefit, choosing very small block sizes such as block sizes of four should be avoided.

Furthermore, it was demonstrated that the variance of the treatment effect estimator is overestimated when using the t-test, which results in very small type I error rates for sequences which are quite well balanced over time. These sequences yield a small bias in the treatment effect estimate, but a much too large variance estimate. Thus, to avoid being overly conservative for time balanced sequences, the blocking has to be taken into account for variance estimation, as recommended by Mats and Lachin [22]. We showed that this is already necessary for small (maybe unobserved) time trends within the two-fold standard deviation. For Efron’s biased coin design [13], it was shown by Rosenkranz [19] that the standard analysis using the t-test can become very anti-conservative. In contrast, he showed that for the truncated binomial design [13] this analysis can become very anti-conservative. For comparison, it should be noted that a larger strength of the time trend was chosen in the paper by Rosenkranz. For response-adaptive randomization procedures, Simon and Simon [23] introduced a randomization test which preserves the nominal significance level.

Due to its standard implementation in every randomization software and its well-known properties, the use of permuted block randomization is very popular in practice. A further important aspect for the use of PBR is that the issue of variance estimation can be easily addressed due to the blocking factor in parametric models. Therefore, we investigated the properties of this randomization procedure. Other randomization procedures [2, 24–27] were developed addressing the issue of balancing properties and predictability. In [18], comparisons of parts of these randomization procedures are presented. Further evaluation and comparison should be the subject of future research, taking into account the concepts presented here.

In addition to the necessity of a good design of clinical trials to reduce chronological bias, an attempt should be made...
to identify time trends [9] and include them in the analysis by use of the consecutive patient number or the recruitment time, as appropriate. However, it cannot be guaranteed that the time trend is identified and, if identified, modeled in an appropriate way (Figure 1).

5. Conclusion and Recommendations

Unobserved time trends can cause a non-negligible bias in the treatment effect estimate and the test decision. Using small block sizes, the maximum extent of bias in the treatment effect estimate and the inflation of the empirical type I error for specific sequences can be controlled. If a serious risk of selection bias exists, we recommend using medium block sizes because they already restrict chronological bias to an acceptable extent. Since the t-test is far too conservative even for small time trends, an analysis of variance which includes the block as a covariate should always be used when the permuted block randomization was applied. To become aware of trends, in the analysis stage of every clinical trial the data should be checked for possible time trends by using the graphical methods suggested by Altman and Royston [9]. Nevertheless, the best way to avoid a strong bias is the choice of a suitable randomization procedure in the design stage of a clinical trial.

Acknowledgment

Part of the work was done within the IDEAl project, which has received funding from the European Union’s 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552.

References