Contents

Chapter 1

Introduction

A clinical trial is an experiment to study the adverse and beneficial effects of a (new or existing) treatment. There are generally four phases of clinical trials. Phase I trials estimate tolerability, pharmacodynamics and pharmacokinetics, and can determine a dose range. Phase II trials determine whether the treatment has any biological effect. Phase III trials asses the effect of the new treatment (or the treatment under consideration) compared the effect of an existing treatment or placebo. Phase IV trials are long term studies after the treatment has been approved [Rosenberger and Lachin, 2002, Friedman et al., 2010].

For Phase III trials, the crucial part is the randomisation of the patients. By randomly assigning patients to one treatment group or another, the comparability among the groups is aided. Several sources of bias are mitigated by using randomisation strategies. Matching groups on known covariates does not guarantee comparability of the groups. Unknown, but possibly important, covariates are not matched among the groups if the matching is done on known covariates. Obviously, matching is not possible on unknown covariates. But, while necessary, randomisation alone is not sufficient to obtain an unbiased study. The outcomes of all patients should be obtained in a similar and unbiased fashion, and any missing data should not bias the comparison of the groups. Even then, severe covariate imbalances can occur in smaller samples [Rosenberger and Lachin, 2002].

Besides mitigating selection bias and accidental bias, randomisation cal also aid inference. Accidental bias is, as described, above, bias introduced by the effect of confounding factors which are not controlled for. Examples are a time trend or a covariate which is ignored. Selection ias is bias introduced by an investigator who tries to 'outsmart' the randomisation procedure and thus tries to influence which patients receive which treatment. Under classical linear model assumptions, as well as under more general statistical models, balanced treatment assignments (an equal number of patients in each treatment arm) is universally optimal: it minimises variance of the estimators and maximises the power of the test. So a random allocation of each patient, as well as a balanced treatment assignment is desirable. When the treatments are forced to be balanced, once one treatment arm is 'full' (contains half the total number of patients available), the rest of the patients are allocated deterministically to the other treatment arm. The other extreme is complete randomisation, where all patients have equal chance to get allocated to either treatment group. This, on the other hand, can lead to severe imbalances. There are many options in between these extremes. In this study, the focus will be on Permuted

Block Randomisation (PBR), Efron's Biased Coin (EBC), and Big Stick Design (BSD). With PBR the allocation is forced to be balanced after each multiple of n patients are allocated, with n the size of the blocks. BSD doesn't enforce exact balance, but instead deterministically allocates subjects to the smaller group when the imbalance between the two treatment arms becomes too large. EBC doesn't enforce balance at all, but instead gives higher probability of allocation to the smaller of the two treatment arms. All three of these procedures give a high probability of reasonably balances groups at any point in the allocation procedure, which is particularly important in phase III clinical trials [Antognini et al., 2015, Rosenberger and Lachin, 2002].

The most commonly used basis for statistical tests is the population model. In the population model it is assumed that the subjects are sampled at random from a reference population. Given the random sampling from this reference population, the responses would then be independent and identically distributed on unknown population parameters. In clinical trials, the subjects are not randomly sampled from some reference population. Instead, patients are recruited in some structured way, and then randomly assigned to one treatment group or another [Rosenberger and Lachin, 2002, Lehmann and D'Abrera, 2006].

Under the randomisation (permutation) model the null hypothesis is that the assigned treatment had no effect on the responses of the patients in the study. So the response that was observed, for each patient, is what would have been observed regardless of the assigned treatment. The way the patients were assigned to the treatment groups, then, caused any observed difference between the treatment groups [Rosenberger and Lachin, 2002, Lehmann and D'Abrera, 2006].

To use randomisation tests under regression models, Parhat et al. [2014], Parhat [2013] use randomisation tests using the residuals of the fitted model as the outcome variable. These can then be used in a rank test, to test differences between the treatment groups. With longitudinal data, using a generalised linear mixed model (GLMM), the residuals are not appropriate. The residuals of a GLMM are not a direct measure of the slope. Instead Parhat et al. [2014], Parhat [2013] suggest to use the predictors of the random slopes from the GLMM and perform a rank test on those.

The study of Parhat et al. [2014], Parhat [2013] compared the size (Type I error probability) and power of the traditional (t-test) inference with their new procedure (using the predictors) for complete randomisation and Efron's Biased Coin, both with complete and conditional reference sets. This was done under correctly specified models and misspecified models. Sample sizes they considered were 60, 96, and 100 [Parhat, 2013, Parhat et al., 2014]. The current study tries to extend the approach to small sample sizes (total sample sizes of 12, and 24), considers two variance-covariance structures (compound symmetry and first order autoregressive), and considers two new randomisation procedures (permuted block randomisation and big stick design).

The structure is as follows. First some background, related to the models and randomisation procedures, is introduced. Then the specific methods used are presented, like the specific parameter values for the simulations and the parameters of the randomisation procedures. Thirdly, the results of the simulations are presented. Finally the results will be discussed.

Chapter 2

Background

2.1 Linear Mixed Models

In this study subjects (patients) are divided into two distinct groups, one receiving treatment A, and the other group receiving treatment B, and each patient is examined multiple times during the course of the study. So there are two levels of clustering: the patients in one treatment group might, as a whole, respond differently to their treatment than the patients in the other group. Secondly, one patient might respond differently to the treatment than another patient with the same treatment. For example, some patients might already have a higher levels of the measured quantity than other patients, and some patients might recover faster (or respond faster) than other patients in the same treatment group. To allow for the individual differences between the responses of the patients, and to still be able to see the differences on the group (treatment) level, linear mixed models can be used. These models have a 'fixed' component, similar to the general linear models, and a 'random' component, which describes the variation in the individual measurements [Verbeke and Molenberghs, 2000].

The linear mixed models in this study have a continuous longitudinal outcome. Y_{ij}, then, is the outcome/measurement at the jth observation/visit of subject i. With a total of N subjects, $i = 1, \ldots, N$, and $j = 1, \ldots, n_i$. In this study, all patients have the same number of measurements, so $j = 1, \ldots, n$. The model itself, in general, can be written as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}) \tag{2.2}$$

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.3}$$

with $\mathbf{Y}_i$ the n_i dimensional vector of responses for subject i, with $i = 1, \ldots, N$ and N the total number of subjects and n_i the number of measurements for subject i. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the matrices of the known covariates, with dimension $(n_i \times p)$ and $(n_i \times q)$, so the number of rows is equal to the number of measurements for the particular subject and the number of columns is equal to the number of fixed and random effects, respectively. $\boldsymbol{\beta}$ and $\mathbf{b}_i$ are the p- and q-dimensional vectors of fixed and random effects, respectively. The fixed effects are the same for all subjects, the random effects are subject-specific. The random effects are normally distributed with mean 0 and symmetric variance-covariance matrix D. The random effects are subject-specific, but the variance-covariance matrix D is the

same for all subjects. The residual components (random errors) are specified by ϵ_i. These are specific for each observation (so can be different for each measurement of a subject), and are independent, identically distributed normal random variables with mean 0 and variance-covariance matrix $\mathbf{\Sigma}$. The random effects and random errors are independent [Lee and Braun, 2012, Verbeke and Molenberghs, 2000, Gelman and Hill, 2007, Ogorek, 2012].

The setting studied here is a clinical trial with two treatments and a continuous longitudinal outcome. The variables in the model are the treatment and the visit (or time). The effect of time is treated as random, as subjects are measured multiple times. The model also has subject-specific (random) intercepts. This model can be specified as

$$Y_{ij} = \beta_1 + b_{1i} + \beta_2 x_{2ij} + (\beta_3 + b_{3i})\, x_{3ij} + \beta_4 x_{2ij} x_{3ij} + \epsilon_{ij} \tag{2.4}$$

with β_1 and b_{1i} the overall intercept and subject-specific deviations from the overall intercept, β_2 and x_{2ij} the effect of the treatment and the assigned treatment, β_3, b_{3i}, and x_{3ij} the overall effect of time, the subject-specific deviations from the overall effect of time, and the time of the measurements, respectively.

The model in equation (2.1) can be rewritten to

$$\mathbf{Y}_i \sim (\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{\Sigma}_i) \tag{2.5}$$

this can be rewritten further as

$$\mathbf{Y}_i \sim (\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i) \tag{2.6}$$

with $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{\Sigma}_i$ [Verbeke and Molenberghs, 2000].

The correlation between the measurements for each subject, and the variability of the measurements can have different structures. In the current study, the repeated measurements are taken over time. Two covariance structures are considered: compound symmetry (CS) and first order autoregressive (AR(1)). In compound symmetry the variance of all measurements for a subject are the same, and the correlation between each of the measurements are also the same. In AR(1) all the variances are still the same, but each measurement depends on the previous measurement for that patient. The variance matrix $\mathbf{V}_i$, with ten measurements per subject, can then be written as

$$V_{jk} = \begin{cases} \sigma^2 & \text{if } j = k \\ \sigma^2\rho & \text{if } j \neq k \end{cases}$$

for compound symmetry, and as $V_{jk} = \sigma^2\rho^{|j-k|}$ for AR(1) [Verbeke and Molenberghs, 2000].

2.2 Type I Error and Power

The Type I error of a test measures how many times the null-hypothesis of no treatment effect is rejected, when the treatment is not effective. In other words, how many times does the test at hand give a false positive result? The power of a test measures how many times the null-hypothesis of no treatment effect is rejected when the treatment is indeed effective. In other words, how many times does the test at hand give a true positive result?

2.2.1 WALD Test

The WALD test, also called the z-test, can be used to test the hypothesis that a single regression parameter is equal to or different from a certain value. In general it tests the, in this case two-sided, hypothesis

$$H_0 : \beta_k = a$$
$$H_1 : \beta_k \neq a$$

with a the hypothesised value of the parameter β_k under the null-hypothesis. The test statistic itself is

$$z^* = \frac{\widehat{\beta}_k - a}{\text{se}\left(\widehat{\beta}_k\right)} \tag{2.7}$$

which follows a standard normal distribution. When it is assumed that under the null-hypothesis there is no effect, the test reduces to

$$z^* = \frac{\widehat{\beta}_k}{\text{se}\left(\widehat{\beta}_k\right)} \tag{2.8}$$

With $\alpha = 0.05$, the null-hypothesis of no effect is rejected if $|z^*| > 1.96$ [Kutner et al., 2005].

The Type I error probability of the WALD test is computed as follows. A large number of data sets are generated under the null hypothesis. A linear mixed model is fit on each of these data sets, and the WALD statistic of β_3 is calculated for each of these. The fraction of test-statistics which are larger than the critical value for a two-sided test at $\alpha = 0.05$ is the Type I error probability of the WALD test. The power of the test is computed in a similar fashion, but the data is generated under the alternative hypothesis: $\beta_3 \neq 0$. The power of the test to find an effect of at least the given size of β_3 is then the fraction of test-statistics which are larger than the critical value for a two-sided test at $\alpha = 0.05$ [Parhat, 2013, Hilgers, 2015].

2.2.2 Randomisation Tests

In parallel two-treatment randomised clinical trials patients are randomly assigned to one of the treatment groups. Different randomisation procedures can give different sets of possible treatment assignments (randomisation sequences). Not all sequences are possible for each randomisation procedure, and not all randomisation sequences from one procedure are equally likely. For example, in balanced randomisation procedures, it would not be possible to have one treatment group larger than the other group. The same sequence of treatment assignments can also have a different probability for different randomisation procedures. In biased coin designs severely unbalanced groups are much less likely than more balanced group-sizes [Rosenberger and Lachin, 2002].

As stated above in the Introduction, the population model for statistical inference does not readily apply to clinical trials. Randomisation provides a basis for a statistical test of the equality of the treatments among the patients enrolled in the study. The basis

for inference using randomisation only relies on the way the patients were assigned to the treatment groups. Under the null hypothesis of no treatment effect, the randomisation sequences have no effect on the outcome variable. The observed responses are instead a set of deterministic values, which are unaffected by the treatment. This means that the arrangement of patient responses only depend on the ways the patients were randomised. The null-distribution of the treatment effect given the randomisation can be calculated exactly, by refitting the model with all of the possible randomisation sequences and weighing the models by the probability of each of the randomisation sequences. The p-value of the observed treatment effect can be calculated as follows: compute the test statistic for each of the possible randomisation sequences, keeping the responses fixed, and sum the probabilities of those randomisation sequences for which the test statistic is larger than the observed test statistic (calculated with the original randomisation sequence) [Rosenberger and Lachin, 2002].

With complete randomisation all possible treatment-assignments are possible, including the cases where all subjects are assigned to the same treatment group. For 2 treatment groups, the number of possible treatment-assignments would be 2^N, which becomes prohibitively large even for small N. With a total sample size of 20 subjects, the number of possible treatment-assignments is already more than 1 million. To overcome this problem, one possible strategy is to randomly sample a subset of all possible randomisation sequences and calculate the Monte Carlo p-value [Rosenberger and Lachin, 2002].

The test statistic used in the randomisation test can be anything with which a difference between two groups can be measured. A test statistic frequently used in randomisation tests is the linear rank test, using with the Van der Waerden scores, of the marginal residuals [Parhat et al., 2014, Parhat, 2013, Lehmann and D'Abrera, 2006, Van der Waerden and Nivergelt, 1956, Hilgers, 2015]. For longitudinal analyses with GLMM this approach does not work, as the residuals are not a direct measure of the slope of the regression line. Instead, Parhat et al. [2014] suggests to use the predicted rate of change from the linear model [Parhat et al., 2014, Hilgers, 2015].

The Type I error rate is found with a two-step procedure. First the critical value of the test is computed, and then the Type I error rate is found. To find the critical value of the test, data is generated under the null-hypothesis of no treatment effect ($\beta_3 = 0$). Then the marginal residuals are computed as

$$Y_{ij} - (\beta_1 + \beta_2 x_{2ij} + b_{3i} x_{3ij} + \beta_4 x_{2ij} x_{3ij}) \tag{2.9}$$

or, in matrix notation

$$Y - X\beta \tag{2.10}$$

Form this, the linear predictors can be calculated as

$$\eta_{ij} = \beta_1 + b_{1i} + \beta_2 x_{2ij} + (\beta_3 + b_{3i})\, x_{3ij} + \beta_4 x_{2ij} x_{3ij} \tag{2.11}$$

These are then ranked, and the Van der Waerden scores calculated as

$$\Psi\left(\frac{r}{N+1}\right) \tag{2.12}$$

with Ψ the inverse normal cumulative distribution function, r the ranks, and N the total sample size [Van der Waerden and Nivergelt, 1956]. If there are less then 5 000 possible

randomisation sequences, all randomisation sequences are computed, otherwise 5 000 possible randomisation sequences are drawn at random. The randomisation sequences are combined with the ranks of the linear predictors. For each of the randomisation sequences the sum of the Van der Waerden scores, S, is calculated for the patients which were assigned to treatment A. The 95% critical value of the test statistic, S_{cr}, is found by taking the 97.5% quantile of these scores S [Parhat et al., 2014, Parhat, 2013, Hilgers, 2015].

The second step is to make a large number (say, 1 000) data sets under the null hypothesis. The first part is similar to the first step: A linear mixed model is fit on each of these data sets, and the linear predictors calculated. These are then ranked, and the Van der Waerden scores calculated. If there are less then 5 000 possible randomisation sequences, all randomisation sequences are computed, otherwise 5 000 possible randomisation sequences are drawn at random. The randomisation sequences are combined with the ranks of the linear predictors. For each of the randomisation sequences the sum of the Van der Waerden scores, S, is calculated for the patients which were assigned to treatment A. For each data set, the proportion of randomisation sequences with $|S| > |S_{cr}|$ is calculated. The mean of these proportions is the Type I error rate of the test. The power of the test is computed in a similar fashion, but the data is generated under the alternative hypothesis: $\beta_3 \neq 0$. The power of the test to find an effect of at least the given size of β_3 is then the fraction of test-statistics which are larger than the critical value [Parhat et al., 2014, Parhat, 2013, Hilgers, 2015].

As not all sequences are possible in a given randomisation strategy, and not all randomisation strategies lead to balanced groups, the question arises what the reference set is for the randomisation test. In other words, which randomisation sequences are used to generate the null distribution of the test. The conditional reference set takes into account only those randomisation sequences, that are possible given the randomisation strategy and that have the same number of patients in treatment A and treatment B as the original randomisation sequence. The unconditional reference set takes into account *all* possible randomisation sequences, given the randomisation strategy and the total number of subjects [Rosenberger and Lachin, 2002]. A comparison of the benefits and drawbacks of both reference sets is beyond the scope of this thesis. In this study only the unconditional reference sets will be used.

2.2.3 Permutation Tests

In the permutation test the linear predictors are permuted among the observations. This is slightly different from the randomisation test, where they were permuted among the randomisation sequences. The procedure to find the Type I error rate for the permutation test is similar to the procedure for the randomisation test. First the critical value for the test statistic is found, and subsequently a large number of data sets under the null hypothesis is generated and the test statistics compared with this critical value.

To find the critical value of the test, data is generated under the null-hypothesis of no treatment effect ($\beta_3 = 0$). Then 5 000 permutations of the response variable are generated. The model of equation (2.4) is fitted to the data sets with the permuted responses, and for each model the linear predictors are computed as in equation (2.11). These are then ranked, and the Van der Waerden scores calculated as in equation (2.12). For each of the

permutations the sum of the Van der Waerden scores, S, is calculated for the patients which were assigned to treatment A. The 95% critical value of the test statistic, S_{cr}, is found by taking the 97.5% quantile of these scores S [Parhat et al., 2014, Amaratunga et al., 2014, Hilgers, 2015].

The second step is to make a large number (say, 1 000) data sets, again under the null hypothesis. The first part is similar to the first step: 5 000 permutations of the response variable are generated. The linear mixed model is fitted to the data sets with the permuted responses, and for each model the linear predictors are computed. These are ranked, and the Van der Waerden scores calculated. For each of the permutations the sum of the Van der Waerden scores, S, is calculated for the patients which were assigned to treatment A. For each data set, the proportion of randomisation sequences with $|S| > |S_{cr}|$ is calculated. The mean of these proportions is the Type I error rate of the test. The power of the test is computed in a similar fashion, but the data is generated under the alternative hypothesis: $\beta_3 \neq 0$. The power of the test to find an effect of at least the given size of β_3 is then the fraction of test-statistics which are larger than the critical value [Parhat et al., 2014, Amaratunga et al., 2014, Hilgers, 2015].

2.3 Randomisation Procedures

2.3.1 Efron's Biased Coin

Efron's Biased Coin design (EBC) is a way to randomise the treatment allocations. Instead of using a 'coin' to allocate each patient to treatment A or B with 50% chance, a biased coin is used to improve the balancing of the treatment arms. Allocating each patient with a 50% chance to treatment A or B gives a non-negligible chance of ending up with (almost) all patients in only one of the two treatment arms. To avoid this, when an imbalance occurs in the treatment allocations, a biased 'coin' is used to reduce the imbalance. When for example 6 patients have been allocated, there is no imbalance if there are 3 patients allocated to each treatment group. It can also happen that more patients have been assigned to one treatment or another (there *has* to be an imbalance if an odd amount of patients have entered the study). If 4 patients have been assigned to treatment A, while only 2 patients have been assigned to treatment B, then there is an imbalance of $4 - 2 = 2$ [Rosenberger and Lachin, 2002].

The procedure for EBC is as follows. The difference in the number of patients assigned to treatment A (n_A) versus the number of patients assigned to treatment B (n_B), after i patients, is the treatment imbalance. This can be written as $D_i = n_A - n_B$. If there are more patients allocated to treatment A, the difference is positive. If there are more patients allocated to treatment B, the difference is negative. If there are an equal number of patients allocated to treatment A and B, the difference is 0. When the i^{th} patient is to be allocated to one of the treatment groups, the treatment imbalance thus far, D_{i-1}, decides how the 'coin' is biased. If $D_{i-1} = 0$, so if the treatment arms are now balanced, the probability of assigning patient i to treatment A is 50%. If $D_{i-1} < 0$, then there are more patients allocated to treatment B than to treatment A, and the probability of assigning patient i to treatment A is p. If $D_{i-1} > 0$, then there are more patients allocated to treatment A than to treatment B, and the probability of assigning patient i to treatment A is $1-p$. With the bias of the coin $p \in (0.5, 1]$ (so $0.5 < p \leq 1$) [Rosenberger

and Lachin, 2002].

The bias of the coin p is the same regardless of the size of the imbalance. The higher the bias of the coin ($p \approx 1$), the smaller the imbalance, but also the more deterministic the treatment assignments will be. The lower the bias of the coin ($p \approx 0.5$), the more random the treatment assignments will be, but the higher the probability of larger imbalances. Antognini et al. [2015] calculated optimal trade-offs between this 'randomness' and 'balance', for several sample sizes. The optimal bias p for small sample size, like considered in this study, was found to be 0.67 [Antognini et al., 2015].

2.3.2 Big Stick Design

Big Stick Design (BSD) is a different way to nudge the treatment assignments into balance. While with EBC the treatments assignments are being nudged into balance immediately when an imbalance occurs, with BSD there is a pre-defined maximum tolerated imbalance (mti). When the imbalance is larger than this mti, the next patient is assigned to the treatment group with the smallest number of patients. If the lowest level of imbalance that is deemed unacceptable is defined as the constant c, and $|D_{i-1}| < c$, so if the treatment arms are now balanced within the accepted tolerance limit, the probability of assigning patient i to treatment A is 50%. If $|D_{i-1}| = c$, the maximum tolerated imbalance is violated and there are too many patients assigned to treatment A relative to treatment B, so the patient is allocated to treatment B. If $|D_{i-1}| = -c$, the maximum tolerated imbalance violated and there are too many patients assigned to treatment B relative to treatment A, so the patient is allocated to treatment A [Rosenberger and Lachin, 2002]. From Hilgers [2015], a suitable value for the maximum tolerated imbalance was found to be 4.

2.3.3 Permuted Block Randomisation

Both EBC and BSD can end with two samples of unequal sizes. To force the samples to have equal size at the end of the treatment allocations, and at various points during the treatment allocations, Permuted Block Randomisation (PBR) can be used. With PBR, patients are not all allocated in one large group, but in several smaller groups. With M blocks, each block contains $m = N/M$ patients, with m and even number. Within each block, an equal number of patients is allocated to treatment A and to treatment B. The allocation of the patients within each block can be done with either the Random Allocation Rule (RAR), or with the truncated binomial design. In this study the focus is on PBR with RAR [for the truncated binomial design see, for example, Rosenberger and Lachin, 2002]. Using the RAR procedure, the probability of assigning the next patient to treatment A is calculated by dividing the one half of the sample size, minus the number of patients already assigned to treatment A, by the total sample size minus sum of the number of patients already assigned to treatment A and treatment B:

$$\frac{\frac{m}{2} - m_A}{m - (m_A + m_B)}$$

In this procedure the 'total sample size' is the size of the blocks m, and the number of patients already assigned to treatment A and treatment B are those within the current block: m_A and m_B [Rosenberger and Lachin, 2002].

Chapter 3

Methods

A series of simulation studies were performed to study the Type I error and the power of the WALD test, randomisation test, and permutation test, under different settings. The simulated model is specified in equation 2.4. 1000 data sets were generated for each of the randomisation strategies. For the randomisation/permutation tests, 1000 randomisations/permutations were performed for each data set.

The settings studied were as follows: The total sample size N was varied between 12 and 24. For each randomisation strategy the WALD test, the randomisation test, and the permutation test was performed. For EBC, the probability of the biased coin to allocate the next patient to the smallest group p was 0.67. The maximum tolerated imbalance for BSD was 4. For PBR block sizes of 4 and 6 were used for simulations using both total sample sizes. For the simulations with total sample size $N = 24$, block sizes of 8 and 12 were also used. This gave a total of ten settings for the randomisation strategies.

The parameter values for the simulated data sets were as follows: The fixed effect for the treatment $\beta_1 = 1$, the fixed effect for time $\beta_2 = 1.5$, and the fixed effect for the interaction between treatment and time $\beta_3 = 0$ for the test for Type I error, and $\beta_3 = 1$ for the test for power. The random effects were sampled from a multivariate normal distribution with mean 0 and variance-covariance matrix for compound symmetry and for AR(1), with the correlation between the measurements of a subject $\rho = 0.1$ and the variance $\sigma^2 \in \{2, 5, 8\}$.

Simulations and model fitting were done using `R` [R Core Team, 2015]. Sampling from the multivariate normal distribution was done using the package `mvtnorm` [Genz et al., 2014]. The GLMMs were fitted using the package `lme4` [Bates et al., 2015]. The WALD test was done using the package `lmerTest` [Kuznetsova et al., 2015]. Schindler [2015] has very kindly provided a beta version of the package `randomizeR` to provide the randomisation sequences.

Table 3.1: Settings and parameter values studied

Variance structure	CS		AR(1)
Variance	2	5	8
Test	WALD Randomisation test Permutation test		
Treatment allocation	EBC	$N = 12$	$p = 0.67$
	BSD	$N = 12$	$mti = 4$
	PBR	$N = 12$	block size = 4
	PBR	$N = 12$	block size = 6
	EBC	$N = 24$	$p = 0.67$
	BSD	$N = 24$	$mti = 4$
	PBR	$N = 24$	block size = 4
	PBR	$N = 24$	block size = 6
	PBR	$N = 24$	block size = 8
	PBR	$N = 24$	block size = 12

Chapter 4

Results

4.1 EBC ($p = 0.67, N = 12$)

The results of the simulations of Efron's Biased Coin, with bias $p = 0.67$, and total sample size $N = 12$, are as follows (Table 4.1).

The power of the WALD test, for both variance structures CS and AR(1), and with variance of 2 and 5, is 1. All the data sets simulated with these parameters, and $\beta_3 = 1$ were found to have an effect for β_3. The power of the WALD test with variance of 8 is 0.999 with data simulated with CS, and 0.994 with data simulated with AR(1). The type I error probability of the WALD test with data simulated with CS is 0.028 for all variances used in this study. This means that in 2.8% of the data sets simulated under these conditions, with $\beta_3 = 0$, a time-varying treatment effect was still found (a false positive finding). The type I error probability of the WALD test with data simulated with AR(1) is 0.036 when the data is simulated with variance 2, and 0.035 when the data is simulated with variance 5 or 8.

The power of the permutation test is 0.051 when the data is simulated with variance 2, 5, or 8 and with AR(1), as well as with CS and variance 2. Of the data sets simulated with these parameters, and $\beta_3 = 1$, 5.1% were found to have an effect for β_3. For data sets simulated with CS and variance 5 or 8, the power is 0.052. The type I error for the randomisation test is 0.051 under all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.286 and 0.287 for CS and AR(1) respectively. With data simulated with variance 5, the power of the randomisation test is 0.203 and 0.201 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.162 and 0.163 for CS and AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.055, 0.052 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.062, 0.055, and 0.052, respectively.

4.2 PBR (block size = 4, $N = 12$)

The results of the simulations of Permuted Block Randomisation, with block sizes equal to 4, and total sample size $N = 12$, are as follows (Table 4.2).

Table 4.1: Simulation results of power and type I error probability for Efron's Biased Coin ($p = 0.67$), with total sample size $N = 12$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.028	1	0.036
	5	1	0.028	1	0.035
	8	0.999	0.028	0.994	0.035
Randomisation	2	0.286	0.055	0.287	0.061
	5	0.203	0.052	0.201	0.055
	8	0.162	0.050	0.163	0.052
Permutation	2	0.051	0.051	0.051	0.051
	5	0.052	0.051	0.051	0.051
	8	0.052	0.051	0.051	0.051

The power of the WALD test, for both CS and AR(1), and with variance of 2 and 5, is 1. The power of the WALD test with variance of 8 is 0.997 with data simulated with CS, and 0.995 with data simulated with AR(1). The type I error probability of the WALD test with data simulated with CS is 0.028 for all variances used in this study. The type I error probability of the WALD test with data simulated with AR(10 is 0.037 when the data is simulated with variance 2, 5, or 8.

The power of the permutation test is 0.052 for data simulated with variance 2, 5, or 8 and with CS and AR(1). The type I error for the randomisation test is 0.051 under all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.294 and 0.295 for CS and AR(1) respectively. With data simulated with variance 5, the power of the randomisation test is 0.200 and 0.195 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.166 and 0.160 for CS and AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.058, 0.052 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.064, 0.058, and 0.054, respectively.

4.3 PBR (block size = 6, $N = 12$)

The results of the simulations of Permuted Block Randomisation, with block sizes equal to 6, and total sample size $N = 12$, are as follows (Table 4.3).

The power of the WALD test, for both CS and AR(1), and with variance of 2 and 5, is 1. The power of the WALD test with variance of 8 is 0.998 with data simulated with CS, and 0.996 with data simulated with AR(1). The type I error probability of the WALD test with data simulated with CS is 0.035 for all variances used in this study. The type I error probability of the WALD test with data simulated with AR(1) is 0.041, 0.039, or

Table 4.2: Simulation results of power and type I error probability for Permuted Block Randomisation with block size = 4, and total sample size $N = 12$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.028	1	0.037
	5	1	0.028	1	0.037
	8	0.997	0.028	0.995	0.037
Randomisation	2	0.294	0.058	0.295	0.064
	5	0.200	0.052	0.195	0.058
	8	0.166	0.050	0.160	0.054
Permutation	2	0.052	0.051	0.052	0.051
	5	0.052	0.051	0.052	0.051
	8	0.052	0.051	0.052	0.051

0.040 when the data is simulated with variance 2, 5, or 8 respectively.

The power of the permutation test is 0.052 for data simulated with variance 2, 5, or 8 and with CS and AR(1). The type I error for the randomisation test is 0.051 under all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.299 for both CS and AR(1). With data simulated with variance 5, the power of the randomisation test is 0.184 and 0.178 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.147 and 0.140 for CS And AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.056, 0.051 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.060, 0.053, and 0.051, respectively.

4.4 BSD ($mti = 4$, $N = 12$)

The results of the simulations of Big Stick Design, with maximum tolerated imbalance equal to 4, and total sample size $N = 12$, are as follows (Table 4.4).

The power of the WALD test, for CS with variances 2 and 5, and AR(1) with variance 2, is 1. The power of the WALD test with variance 8 for CS, and with variance 5 for AR(1), is 0.999. The power for data simulated with variance 8 and AR(1) is 0.993. The type I error probability of the WALD test with data simulated with CS is 0.035 for all variances used in this study. The type I error probability of the WALD test with data simulated with AR(1) is 0.040 when data is simulated with variance 2, and 0.041 when the data is simulated with variance 5, or 8.

The power of the permutation test is 0.051 for data simulated with variance 2, 5, or 8 with CS, and with variance 2 or 5 with AR(1). The power of the permutation test with variance 8 and AR(1) is 0052. The type I error for the randomisation test is 0.051 under

Table 4.3: Simulation results of power and type I error probability for Permuted Block Randomisation with block size = 6, and total sample size $N = 12$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.035	1	0.041
	5	1	0.035	1	0.039
	8	0.998	0.035	0.996	0.040
Randomisation	2	0.299	0.056	0.299	0.060
	5	0.184	0.051	0.178	0.053
	8	0.147	0.050	0.140	0.051
Permutation	2	0.052	0.051	0.052	0.051
	5	0.052	0.051	0.052	0.051
	8	0.052	0.051	0.052	0.051

all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 and CS is 0.285. For data simulated with variance 2 and AR(1), the power is 0.284. With data simulated with variance 5, the power of the randomisation test is 0.200 and 0.201 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.160 and 0.161 for CS And AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.056, 0.050 and 0.049, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.063, 0.056, and 0.052, respectively.

4.5 EBC ($p = 4$, $N = 24$)

The results of the simulations of Efron's Biased Coin, with bias $p = 4$, and total sample size $N = 24$, are as follows (Table 4.5).

The power of the WALD test is 1 for all conditions studied. The type I error probability of the WALD test with data simulated with CS is 0.040 for all variances used in this study. The type I error probability of the WALD test with data simulated with AR(1) is 0.046 when data is simulated with variance 2 or 8, and 0.045 when the data is simulated with variance 5.

The power of the permutation test is 0.051 for all conditions studied. The type I error for the randomisation test is 0.050 under all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.266, for both CS and AR(1). With data simulated with variance 5, the power of the randomisation test is 0.191 and 0.189 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.152 and 0.150 for CS And AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS

Table 4.4: Simulation results of power and type I error probability for Big Stick Design with maximum tolerated imbalance = 4, and total sample size $N = 12$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.035	1	0.040
	5	1	0.035	0.999	0.041
	8	0.999	0.035	0.993	0.041
Randomisation	2	0.285	0.056	0.284	0.063
	5	0.200	0.050	0.201	0.056
	8	0.160	0.049	0.161	0.052
Permutation	2	0.051	0.051	0.051	0.051
	5	0.051	0.051	0.051	0.051
	8	0.051	0.051	0.052	0.051

and with variance 2, 5, and 8 is 0.054, 0.051 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.057, 0.056, and 0.053, respectively.

4.6 PBR (block size = 4, $N = 24$)

The results of the simulations of Permuted Block Randomisation, with block sizes equal to 4, and total sample size $N = 24$, are as follows (Table 4.6).

The power of the WALD test is 1 for all conditions studied. The type I error probability of the WALD test with data simulated with CS is 0.038, 0.036, and 0.040 for variance 2, 5, and 8 respectively. The type I error probability of the WALD test with data simulated with AR(1) is 0.042 when data is simulated with variance 2 or 5, and 0.043 when the data is simulated with variance 8.

The power of the permutation test is 0.052 for all conditions studied. The type I error for the randomisation test is 0.051 under all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.275, for CS, and 0.274 for AR(1). With data simulated with variance 5, the power of the randomisation test is 0.213 and 0.214 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.168 for both CS and AR(1). The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.053, 0.051 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.057, 0.056, and 0.053, respectively.

Table 4.5: Simulation results of power and type I error probability for Efron's Biased Coin ($p = 0.67$), with total sample size $N = 24$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.040	1	0.046
	5	1	0.040	1	0.045
	8	1	0.040	1	0.046
Randomisation	2	0.266	0.054	0.266	0.057
	5	0.191	0.051	0.189	0.056
	8	0.152	0.050	0.150	0.053
Permutation	2	0.051	0.050	0.051	0.050
	5	0.051	0.050	0.051	0.050
	8	0.051	0.050	0.051	0.050

Table 4.6: Simulation results of power and type I error probability for Permuted Block Randomisation, with block size = 4, and total sample size $N = 24$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.038	1	0.042
	5	1	0.036	1	0.042
	8	1	0.037	1	0.043
Randomisation	2	0.275	0.053	0.274	0.057
	5	0.213	0.051	0.214	0.056
	8	0.168	0.050	0.168	0.053
Permutation	2	0.052	0.051	0.052	0.051
	5	0.052	0.051	0.052	0.051
	8	0.052	0.051	0.052	0.051

Table 4.7: Simulation results of power and type I error probability for Permuted Block Randomisation, with block size = 6, and total sample size $N = 24$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.024	1	0.034
	5	1	0.024	1	0.034
	8	1	0.024	1	0.034
Randomisation	2	0.263	0.053	0.262	0.057
	5	0.207	0.051	0.210	0.055
	8	0.165	0.050	0.167	0.052
Permutation	2	0.052	0.051	0.051	0.051
	5	0.051	0.051	0.052	0.051
	8	0.051	0.051	0.051	0.051

4.7 PBR (block size = 6, $N = 24$)

The results of the simulations of Permuted Block Randomisation, with block sizes equal to 6, and total sample size $N = 24$, are as follows (Table 4.7).

The power of the WALD test is 1 for all conditions studied. The type I error probability of the WALD test with data simulated with CS is 0.023 for all variances used in the study. The type I error probability of the WALD test with data simulated with AR(1) is 0.034 for all variances used in the study.

The power of the permutation test is 0.052 for data simulated with CS and variance 2, and with AR(1) and variance 5. The power is 0.051 for all other conditions studied. The type I error for the randomisation test is 0.051 under all conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.263, for CS, and 0.262 for AR(1). With data simulated with variance 5, the power of the randomisation test is 0.207 and 0.210 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.165 and 0.167 for CS and AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.053, 0.051 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.057, 0.055, and 0.052, respectively.

4.8 PBR (block size = 8, $N = 24$)

The results of the simulations of Permuted Block Randomisation, with block sizes equal to 8, and total sample size $N = 24$, are as follows (Table 4.8).

The power of the WALD test is 1 for all conditions studied. The type I error probability of the WALD test is 0.033 for all conditions studied.

The power of the permutation test is 0.051 for data simulated with CS and variance 2 or 5, and with AR(1) and variance 2. The power is 0.052 for the other conditions studied.

Table 4.8: Simulation results of power and type I error probability for Permuted Block Randomisation, with block size = 8, and total sample size $N = 24$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.033	1	0.033
	5	1	0.033	1	0.033
	8	1	0.033	1	0.033
Randomisation	2	0.257	0.054	0.256	0.059
	5	0.202	0.052	0.205	0.057
	8	0.163	0.050	0.166	0.055
Permutation	2	0.051	0.050	0.051	0.051
	5	0.051	0.050	0.052	0.050
	8	0.052	0.050	0.052	0.050

The type I error for the randomisation test is 0.051 for data simulated with AR(1) and variance 2, and 0.050 for the other conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.257, for CS, and 0.256 for AR(1). With data simulated with variance 5, the power of the randomisation test is 0.202 and 0.205 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.163 and 0.166 for CS and AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.054, 0.052 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.059, 0.057, and 0.055, respectively.

4.9 PBR (block size = 12, $N = 24$)

The results of the simulations of Permuted Block Randomisation, with block sizes equal to 12, and total sample size $N = 24$, are as follows (Table 4.9).

The power of the WALD test is 1 for all conditions studied. The type I error probability of the WALD test is 0.026 for all conditions studied with CS. The type I error probability of the WALD test with data simulated with AR(1) and variance 2, 5, and 8 is 0.032, 0.031, and 0.033 respectively.

The power of the permutation test is 0.051 for all conditions studied. The type I error for the randomisation test is 0.051 for data simulated with AR(1) and variance 5 and 8, and 0.050 for the other conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.252, for CS, and 0.247 for AR(1). With data simulated with variance 5, the power of the randomisation test is 0.196 and 0.200 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.160 and 0.162 for CS and AR(1) respectively. The type I error probability of the randomisation test with data simulated

Table 4.9: Simulation results of power and type I error probability for Permuted Block Randomisation, with block size = 12, and total sample size $N = 24$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.026	1	0.032
	5	1	0.026	1	0.031
	8	1	0.026	1	0.033
Randomisation	2	0.252	0.055	0.247	0.059
	5	0.196	0.051	0.200	0.056
	8	0.160	0.050	0.162	0.053
Permutation	2	0.051	0.050	0.051	0.050
	5	0.051	0.050	0.051	0.051
	8	0.051	0.050	0.051	0.051

with CS and with variance 2, 5, and 8 is 0.055, 0.051 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.059, 0.056, and 0.053, respectively.

4.10 BSD ($mti = 4$, $N = 24$)

The results of the simulations of Big Stick Design, with maximum tolerated imbalance equal to 4, and total sample size $N = 24$, are as follows (Table 4.10).

The power of the WALD test is 1 for all conditions studied. The type I error probability of the WALD test is 0.027 for CS with variance 2 and 8, and 0.028 for CS with variance 5. The type I error probability of the WALD test with data simulated with AR(1) is 0.032, for all variances studied.

The power of the permutation test is 0.051 for all conditions studied. The type I error for the randomisation test is 0.051 for data simulated with CS and variance 5, and 0.050 for the other conditions that were studied.

The power of the randomisation test, with data simulated with variance 2 is 0.267 for both CS and AR(1). With data simulated with variance 5, the power of the randomisation test is 0.192 and 0.189 for CS and AR(1) respectively. With data simulated with variance 8, the power of the randomisation test is 0.153 and 0.151 for CS and AR(1) respectively. The type I error probability of the randomisation test with data simulated with CS and with variance 2, 5, and 8 is 0.055, 0.051 and 0.050, respectively. The type I error probability of the randomisation test with data simulated with AR(1) and with variance 2, 5, and 8 is 0.059, 0.056, and 0.053, respectively.

Table 4.10: Simulation results of power and type I error probability for Big Stick Design, with maximum tolerated imbalance = 4, and total sample size $N = 24$

	variance	**CS**		**AR(1)**	
		power	*type I*	*power*	*type I*
WALD	2	1	0.027	1	0.032
	5	1	0.028	1	0.032
	8	1	0.027	1	0.032
Randomisation	2	0.267	0.054	0.267	0.058
	5	0.192	0.051	0.189	0.055
	8	0.153	0.050	0.151	0.053
Permutation	2	0.051	0.050	0.051	0.050
	5	0.051	0.051	0.051	0.050
	8	0.051	0.050	0.051	0.050

Chapter 5

Discussion and Conclusions

The aim of this study was to compare the randomisation test with the WALD and permutation test, on type I error probability and power, for small sample sizes in the setting of GLMMs in clinical trials. Ten randomisation strategies were used, and data was simulated with two variance structures and three values for the variance. The variance structures considered were compound symmetry and first order autoregressive. The variances used in the simulations were 2, 5, and 8. Data sets with two sample sizes were generated: with total sample size $N = 12$, and $N = 24$ The randomisation strategies used were Efron's Biased Coin (with bias $p = 0.67$), Permuted Block Randomisation, and Big Stick Design (with maximum tolerated imbalance equal to 4). The block sizes for PBR were 4 and 6 for $N = 12$, and 4, 6, 8, and 12 for $N = 24$.

In general, comparing the randomisation test to the WALD test, it seems that for all conditions the power of the WALD test is higher and the type I error probability of the WALD test is lower. The randomisation test seems to have higher power than the permutation test in all conditions that were considered, but also seems to have higher type I error probability under most conditions. For CS with variance 5 or 8, and for AR(1) with variance 8, the type I error of the randomisation test seems to be similar to the type I error of the permutation test (around 0.050). For the other conditions, the type I error of the randomisation test seems to be slightly higher for the randomisation test (near or around 0.06).

The results from the data simulated with CS can be compared to the results from the data simulated with AR(1). There seems to be no difference between the power and type I error for the permutation test. For the WALD test there seems to be no difference between the power when the data is simulated under CS or under AR(1). The type I error probability of the WALD test seems to be slightly higher when the data is generated under AR(1), relative to when the data is generated under CS with equal variance. The only exception seems to be PBR with block size 8 and $N = 24$, where there seems to be no difference between the results from data sets simulated with CS and AR(1). For the randomisation test, the type I error probability when the data sets are generated with AR(1) seems to be slightly higher than when the data sets are generated with CS. The power of the randomisation test seems to be fairly equal between the data sets generated with CS and AR(1).

For both the WALD test and the permutation test, there seems to be no difference in power and type I error probability when the variance in the simulated data is changed.

For the randomisation test, when the data sets are generated with larger variance, but the variance structure and randomisation procedure are kept equal, both the power and type I error probability seem to go down.

Focusing on the three randomisation strategies individually, the following can be seen. With the WALD test, EBC with total sample size 12 seems to have lower type I error than EBC with total sample size 24, but similar power. With the randomisation test, EBC with total sample size 12 seems to have similar type I error probability relative to EBC with total sample size 24, but lower power. For BSD, the WALD test with total sample size 24 seems to have a lower type I error probability relative to BSD with total sample size 12, but similar power. With the randomisation test, BSD with total sample size 24 seems to have similar type I error probability relative to BSD with total sample size 12, but lower power. For PBR, while there seem to be differences in power and type I error probability for the tests between the sample sizes and between the block sizes, there seems to be no structure in these differences.

The three randomisation sequences can also be compared with each other. For the WALD test, the power under the different randomisation strategies seems similar. With total sample size 12, the WALD test seems to have a lower type I error probability under EBC than with BSD, with PBR in between the two. With total sample size 24, the WALD test seems to have a higher type I error probability under EBC than with BSD, with PBR again in between the two. For the randomisation test, there don't seem to be big differences between the three randomisation strategies. With total sample size 24, PBR with smaller blocks seems to have slightly higher power than PBR with larger blocks, and similar type I error probability.

Parhat et al. [2014], Parhat [2013] developed the specific test statistic used in the randomisation test which was used in this study. The 'old' test statistic uses the ranks of the residuals, whereas this new test statistic uses the linear predictors of the random slopes from the GLMM. As the residuals themselves are not a measure of the slopes of the regression lines, the 'old' procedure could not be used in GLMMs with longitudinal data. The new procedure solves this problem by using the linear predictors instead [Parhat et al., 2014, Parhat, 2013].

In their development of the new procedure Parhat et al. [2014] compared the Type I error probability of the new procedure against the traditional inference (using the t-test), under several conditions. Parhat et al. [2014] and Parhat [2013] used both the complete and conditional reference set for the randomisation test, looked at the effects of misspecifying several parts of the models (but only misspecifying one part of the model at a time), and used total sample sizes of 60, 96 and 100. The randomisation procedures used by them were complete randomisation and Efron's Biased Coin. The bias for EBC used in their study is also $p = 0.67$ [Parhat et al., 2014, Parhat, 2013].

The current study expanded on Parhat et al. [2014] in three ways. Firstly, more randomisation procedures were used. EBC was used again, as in Parhat et al. [2014], with the same bias. While complete randomisation was not used in the current study, two others were: BSD (with maximum tolerated imbalance equal to 4) and PBR (with block sizes 4, 6 for $N = 12$, and 4, 6, 8, and 12 for $N = 24$). Secondly, this study compared the tests with smaller sample sizes. The total sample sizes used in this study were 12 and 24. The smallest sample size used by Parhat et al. [2014] and Parhat [2013] is already more than twice the sample size of the largest sample size considered in this study. As

phase III clinical trials do not always have samples as large as those considered by Parhat et al. [2014] and Parhat [2013] [Rosenberger and Lachin, 2002], it is important to look at what happens to the test statistics under small sample sizes. Thirdly, in this study the randomisation test is also compared to the permutation test.

Where Parhat et al. [2014] and Parhat [2013] also consider the effects of model misspecifications, this study does not. To study the model misspecifications, the data could be generated differently from the distributional assumptions.

Another thing that could be studied is the effect of different parameter values. In this study all parameter values were fixed, apart from the several values for the variance and the variance structure. In addition to this, power and Type I error could be studied under different values for the fixed effects. Another interesting aspect would be to study the effect of multiple random effects. In a clinical trials setting this could, for example, translate to a random effect for clinics.

Bibliography

Dhammika Amaratunga, Javier Cabrera, and Ziv Shkedy. *Exploration and Analysis of DNA Microarray and Other High Dimensional Data*. Wiley, 2nd edition, 2014.

Alessandro Baldi Antognini, William F. Rosenberger, Yang Wang, and Maroussa Zagoraiou. Exact optimum coin bias in Efron's randomization procedure. *Statistics in Medicine*, 2015. ISSN 1097-0258. doi: 10.1002/sim.6576. URL `http://dx.doi.org/10.1002/sim.6576`.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2015. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-8.

Lawrence M. Friedman, Curt D. Furberg, and David L. DeMets. *Fundamentals of Clinical Trials*. Springer, 4th edition, 2010.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2007.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2014. URL `http://CRAN.R-project.org/package=mvtnorm`. R package version 1.0-2.

Ralf-Dieter Hilgers. Outline. personal communication, 2015.

Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill, 5th edition, 2005.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models*, 2015. URL `http://CRAN.R-project.org/package=lmerTest`. R package version 2.0-29.

Oliver E. Lee and Thomas M. Braun. Permutation tests for random effects in linear mixed models. *Biometrics*, 68(2):486–493, 2012.

E.L. Lehmann and H.J.M. D'Abrera. *Nonparametrics: statistical methods based on ranks*. Springer, revised first edition, 2006.

Ben Ogorek. Hierarchical linear models and lmer. *Anything but R-bitrary*. `http://anythingbutrbitrary.blogspot.com/2012/10/hierarchical-linear-models-and-lmer.html`, 2012. Accessed: 2015-07-04.

Parwen Parhat. *Randomization Tests for Regression Models in Clinical Trials*. PhD thesis, George Mason University, Fairfax, VA, 2013.

Parwen Parhat, William F. Rosenberger, and Guoqing Diao. Conditional Monte Carlo randomization tests for regression models. *Statistics in Medicine*, 33(18):3078 – 3088, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

William F. Rosenberger and John M. Lachin. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons, 2002.

David Schindler. *randomizeR: Randomization - Assessment and Evaluation of Randomization Procedures*, 2015. R package version 0.1.

B.L. Van der Waerden and E. Nivergelt. *Tafeln zum Vergleich zweier Stichproben mittels X-Test und Zeichen-test (Tables for comparing two samples by X-test and sign test)*. Springer, 1956.

Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. Springer, 2000.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Type I Error and Power of Randomisation Tests using Linear Mixed Models in the Setting of Small Clinical Trials

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2015**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Burger, Bram

Datum: **28/08/2015**