Sample size estimation for RCTs with repeated measures designs

Andrew Saul

School of Mathematics and Statistics
University of Sheffield

The University Of Sheffield.

Dissertation submitted as part of the requirements for the award of MSc in Statistics, University of Sheffield, 2019–2020

Acknowledgements

I would like to thank my supervisor Professor Stephen Walters for his assistance and guidance throughout this project. I thank my friend and mentor Dr Jim Kay for being willing and available for the last 8 years of my statistical educational journey. Finally, I thank my family Sarveshni, Shreya and Shivan who have been supportive, loving and encouraging during the many highs and lows.

Abstract

Sample size estimation is an essential step in designing a clinical trial. In repeated measures randomized clinical trial designs, measurements observed for each patient are most likely correlated. Specific techniques are required to account for such correlations when estimating sample size. A literature review revealed that linear mixed models (LMM) and generalised estimating equations were methods found to be widely used to compare differences between treatment groups in parallel-group design RCTs using repeated measures. To estimate sample size of such a study modeled with LMM, the estimated model parameters obtained from analyzing a similar study were determined. Power analysis was performed by generating simulated datasets using the LMM and undertaking hypothesis testing using nested models. Factors examined that affected statistical power included: the type of model utilized for hypothesis testing, sample size, correlation, number of time points at which patients' observations were made, and the pattern of missing data. Manipulation of these factors resulted in alterations of study power that were in line with expectations based on the current literature. For this study, the estimated power was in good agreement with the relevant sample size formula that is available in the literature.

Abbreviations

- AR – autoregressive (of order 1)
- AUC – area under the curve
- CRD - clinically relevant difference
- cRCT - cluster randomised controlled trial
- GEE - generalised estimating equations
- GLRT - generalised likelihood ratio test
- ICC - intraclass correlation
- LMM - linear mixed-model
- MAR - missing at random
- MCAR – missing completely at random
- ML - maximum likelihood
- QoL – quality of life
- RCT - randomised controlled trial
- REML - restricted maximum likelihood
- rmANOVA - repeated measures analysis of variance
- TAD – time averaged difference

Contents

Acknowledgements **iii**

Abstract **v**

Abbreviations **vii**

1 Introduction **1**
1.1 Sample size, RCT and repeated measures 1
1.2 Repeated measures trial size considerations 2
1.3 Methods of sample size estimation 3
1.4 Aims 3
1.5 Outline 4

2 Literature Review **5**
2.1 Summary methods 5
2.2 Model based methods 6
2.2.1 rmANOVA 6
2.3 Repeated measures cluster RCTs (cRCTs) 7
2.4 Factors for investigation 7

3 LMM and GEE **9**
3.1 Linear Model 9
3.2 Linear Mixed Model (LMM) 9
3.2.1 Random intercept model 10
3.2.2 Random slope model 11
3.3 LMM matrix notation 11
3.4 Sample correlation 12
3.5 Marginal GEE models 12
3.5.1 Correlation structure 13
3.5.2 Quasi-Likelihood 14
3.5.3 Parameter Estimates 14
3.6 Sample size formulas 15
3.7 Comparison between LMM and GEE 16
3.8 Review of repeated measures cRCT 17
3.8.1 Definition 17

3.8.2 Purpose . . . 17
3.8.3 Analysis . . . 18
3.8.4 Sample size estimation . . . 18

4 Method development for power analysis 19
4.1 Introduction . . . 19
4.2 Study data . . . 19
4.2.1 Outcome measure and study design . . . 19
4.2.2 Sample size . . . 20
4.2.3 Exploratory Data Analysis . . . 20
4.3 The statistical model . . . 22
4.3.1 Choice of Model . . . 22
4.3.2 Explanatory variables . . . 22
4.3.3 Variable re-scaling . . . 23
4.3.4 Inclusion of explanatory effects . . . 23
4.4 Model checking . . . 24
4.5 Simulations . . . 24
4.5.1 Power analysis . . . 24

5 Results and Discussion 27
5.1 Introduction . . . 27
5.2 Type I error . . . 27
5.3 Power curves at ICC = 0.5 . . . 31
5.4 Low sample size . . . 34
5.5 Correlation . . . 35
5.6 Independent observations . . . 36
5.7 Number of sampled time points . . . 38
5.8 Missing data . . . 42
5.9 Covariance matrix misspecification . . . 47
5.10 Limitations and recommendations . . . 47

6 Conclusions 49

A R Code for Model Specification 51

B Model checking 57

C Simulation R code 61
C.1 Complete cases data simulations . . . 61
C.1.1 Power function . . . 61
C.1.2 Calculating power curve . . . 64
C.2 Missing data simulations . . . 65
C.2.1 Power function . . . 65
C.2.2 Calculating the power curve for missing data . . . 67

References 69

Chapter 1

Introduction

1.1 Sample size, RCT and repeated measures

Sample size estimation is an essential step in planning a clinical trial (Julious and A. 2010, p2). When a study is based on an inadequate sample size, the probability of detecting a true effect is reduced, thereby reducing the chances of answering the question posed. On the other hand, if a study utilises a greater sample size than required, resources would be wasted. In addition, the study duration would likely increase and patients may be subjected inferior standard treatments for a longer period than necessary. It can be argued that in both aforementioned scenarios it is unethical to perform a clinical trial. It is therefore important to determine in most cases (excluding adaptive designs) the required sample size of a study before commencement of patient recruitment.

A randomised controlled trial (RCT) is performed by randomly allocating patients into one of two or more treatment groups, where one treatment group acts as a control. The act of random patient allocation reduces systemic differences between groups. Most RCTs aim to determine if one interventional treatment is superior to another, usually a standard treatment (Piaggio et al. 2012) . This type of study is known as a superiority trial. Superiority trials are analysed utilising hypothesis testing. The null hypothesis assumes that no difference of a parameter exists between the two treatments, whereas the alternative hypothesis proposes that a difference of that parameter exits between to the interventions. Possible parameters include the mean outcome response or the rate of change of the outcome response with time. An array of statistical tests can be employed to determine whether to accept or reject the null hypothesis. The body of this thesis will focus on superiority trials between an interventional treatment and a control or standard treatment. It is recognised that in many RCTs participants are not always referred to as patients but subjects, and not all treatments are actually interventional treatments.

In the context of this report a repeated measures (or longitudinal) design refers to a patient receiving a signal treatment, rather than several treatments (crossover

design). Unlike a cross-sectional study, where a single measurement is acquired per patient , in a repeated measures design multiple measurements are obtained. Often the a pre-randomised baseline is initially acquired per patient, followed by one or more post-randomisation measurements. The measurements obtained for a single patient are however most likely correlated. This correlation must be taken into account for accurate sample size estimation.

1.2 Repeated measures trial size considerations

In order to discover with a high probability if a clinically relevant difference between two treatments exist, the estimated sample size required for repeated measures is dependent on the following specifications (Guo et al. 2013):

- Trial Objective
- Primary outcome
- Method of analysis
- Variability of outcome measure
- correlation between measurements for each patient
- number of measurements acquired for each patient
- Clinically relevant difference (CRD)
- Type I error rate
- Power

Examples of trial objectives include non-inferiority, equivalence and bioequivalence (Julious and A. 2010, p27). This report will focus on the superiority of an interventional treatment over a standard treatment. The primary outcome of a trial should allow assessment of the trial objective. Data that consists of continuous primary outcomes will be the focus of this report.

Analysis of repeated measures studies is reviewed in chapter 2 and 3.

Accurate estimation of the unexplained variability in the outcome response (σ^2) and the correlation (ρ) among repeated measures can be challenging. Several strategies have been proposed to select variance and correlation parameters (Guo et al. 2013). These include : (1) utilising data from previous studies, (2) utilising data from pilot studies or (3) making an educated guess from previous experience.

The CRD is considered the smallest difference between the mean response d_m or between the rate of change (slope) of the mean response d_s for the standard and interventional treatment groups that is worth detecting from a clinical perspective. Any smaller difference detected between the two groups is assumed to have no clinical relevance. Statistical testing is based on two-sided tests. The assignment of the CRD is a clinical and not a statistical decision.

The type I error rate (α) represents the probability of incorrectly rejecting the null hypothesis that no difference in the mean response or rate of change of the mean response between standard and the interventional treatment groups exists. The rate is commonly set at a value $\alpha = 0.05$. The alternative hypothesis is that

an increase or decrease in mean response or rate of change of the mean response in the interventional treatment group compared to the standard treatment group. The statistical test is two-sided and the null hypothesis will be rejected if the test statistic is greater than the 97.5% quantile or less than the 2.5% quantile of the probability distribution.

The power $(1-\beta)$ of a study is defined as the probability of a study to reject the null hypothesis when it is false. The term β represents the associated (type II) error rate. Power can be described as the probability of detecting the effect of interest, given that one exists. The value of the study power is commonly set at a level of 0.8 or 0.9. The power of a study increases as the sample size increases, and sample size can be presented in terms of study power. For example, the general aim of power analysis is to determine the power of a study design, or the sample size required to reach a particular power (Johnson et al. 2015).

1.3 Methods of sample size estimation

In this report sample size estimation techniques are considered for a two group superiority RCT, where each group contains the same number of patients. Sample size estimation can be performed utilising closed form analytical calculations or through simulations. Closed formulas usually provide a quick and easy method to determine sample size requirements, but are usually approximations. There may not be enough information available to perform precise calculations. In addition, sample size calculations are very sensitive to assumptions. Sample size calculations may not account for the required pattern of missing data nor the inclusion of covariates, both of which affect sample size estimation (Guo et al. 2013). In addition, formulas are often not available for complex experimental designs (Petras 2016). The potential precision lost using a formula for sample size estimation can be overcome by utilising Monte Carlo simulations (Johnson et al. 2015).

1.4 Aims

This report has the following aims:

- To review the literature of the methods of sample size estimation and analysis in RCTs using repeated measures designs
- To extract parameters from a repeated measures dataset in order to simulate new datasets
- To consider the manner in which the power of hypothesis tests, defined by the comparison of two nested statistical models, is related to various factors, such as sample size, model selection, CRD, correlation and missing data patterns.

1.5 Outline

This report is comprised of five chapters. A literature review of sample size estimation in RCTs for repeated measures designs is contained in chapter 2. A theoretical analysis of the common models used in the field, and associated sample size formulas are discussed in chapter 3. The methods utilised to extract parameters from a published dataset are discussed in chapter 4. Within this chapter the method for generating simulated datasets using these parameters is elucidated. Furthermore, hypothesis testing (the method of power determination) of nested models generated from the simulated datasets is explained. A description of the parameters within models that affect power determination and that will be explored in this report are listed. Chapter 5 describes and discusses the effects of parameters investigated in this report on power. Limitations and recommendations from this report are contained in chapter 5. Chapter 6 contains the conclusions of this report. Appendix A contains R code for the determination of model parameters utilised to model the data of Thomas et al. (2006). Appendix B contains a reporting checking the assumptions of this derived model. Appendix C contains the R code for simulating datasets and performing power analysis with and without missing data.

Chapter 2

Literature Review

A literature review was performed to assess methods of sample size estimation for RCTs using repeated measures designs. A search was performed using the database SCOPUS with the keywords *sample size estimation* and *RCT* and " *repeated measure* or *longitudinal data* or *longitudinal design* ". As no documents were found, the search was repeated without the term *RCT* . A total of 23 articles were found. Further articles were found from citations within these 23 articles.

Sample size estimation methods were found to be dependent on the manner in which repeated measurement data were analysed. Some methods involved aggregating the data into one or two summary measures. Others involved modeling the data. A review of each is provided below.

2.1 Summary methods

In many clinical trials the main goal is to assess the average response to treatment over time, where the response often occurs quickly and remains relatively steady over time. Frison and Pocock (1992) described three ways of analysing longitudinal data that involved data aggregation. The first is termed *POST*, and involves using the mean of each patient's post-treatment measurements as the summary measure. The second is termed *CHANGE* and involves subtraction of the mean baseline measurement(s) from the mean post-treatment measurements. For these two methods a difference in summary measure between the two treatment groups can be tested using a two sample two-sided t-test. The third methods is termed *ANCOVA* (analysis of covariance), where the mean baseline measurement for each patient is utilised as a covariate in a linear model for a comparison of post-treatment means (Frison and Pocock 1992). A covariate-adjusted difference in means between the two treatment groups can be performed. Of the three methods described above, Frison and Pocock (1992) recommend the use of *ANCOVA*.

Another common summary method involves calculating the area under the curve (AUC) (Matthews et al. 1990). A response curve is constructed for each patient and the single area under the curve measure is calculated. As described above, a two sample two-sided t-test can be performed to determine if a difference in AUC can be detected between the two treatment groups.

Summary measures are generally simple to interpret. If missing values are present, summary measures can usually still be produced (Matthews et al. 1990), although biases may be introduced. By summarising data, dependency information (correlation) between measurements within each patient is lost. This leads to an increase in total variation of the sample population and results in an increased sample size requirement to detect a difference between treatment groups.

2.2 Model based methods

Methods that aim to model the longitudinal nature of the data and make inferences about the regression parameters of primary interest are based on recognising the likely correlation structure in the data. Two types of modeling methods were identified in the literature. The first is based on building an explicit parametric model of the data correlation structure. Both *repeated measures analysis of variance* (rmANOVA) (Overall and Starbuck 1979; Overall and Doyle 1994) and *linear mixed-models* (LMM) (Diggle et al. 2002, chap. 9) are based on this parametric model of the correlation structure. The second type is based on a marginal model analysed using *generalised estimating equations* (GEE), whereby the regression of the response variable on the explanatory variables is modeled separately from the within-subject correlation (Liang and Zeger 1986; Liu and Liang 1997). An explanation of LMM and marginal models is provided in chapter 3. These two models have gained more widespread use than rmANOVA, for reasons that will be now described. For the sake of clarity, marginal models analysed with GEE will be referred to as "GEE models".

2.2.1 rmANOVA

rmANOVA is an analysis technique for repeated measure designs. Data is modeled in a similar way as to the agricultural ANOVA designs (Tango 2017, chap. 3). Treatments are considered as main effects (main plots) and subjects are represented as units of the main effects (subplots). In ANOVA, allocation of all units at all plot levels is random. However, in rmANOVA, time units for each subject cannot be randomly allocated. As an result, an unobserved random effect is introduced into the rmANOVA model to overcome this problem. In this sense, rmANOVA has a random effects structure similar to linear mixed models.

Two major drawbacks exist with rmANOVA. rmANOVA can only be performed on complete cases. While imputation methods are available to accommodate missing data, large amounts of missing data can seriously compromise the analysis

(Diggle et al. 2002, p125). In addition, rmANOVA fails to exploit potential gains in efficiency in modeling the covariance among repeated measurements (Diggle et al. 2002, p114). Both LMM and GEE methods do not have these disadvantages inherent in rmANOVA.

2.3 Repeated measures cluster RCTs (cRCTs)

Two articles were found during the literature review that were based on repeated measures cluster RCT designs (Tu et al. 2006; Guthrie et al. 2012). While this report does not focus on cRCTs, a brief discussion is provided in section 3.8.

2.4 Factors for investigation

A majority of articles found in the Scopus search focused on deriving sample size formulas for GEE models. One topic of interest investigated in many articles was the effect of missing data (Wang, Zhang, and Ahn 2020; Lou et al. 2017a, 2017b; Lou, Cao, and Ahn 2017; Liu and Liang 1997; Ahn and Jung 2005). Closed formulas for these GEE models incorporating a variety of missing data patterns were derived and their accuracy assessed using Monte-Carlo simulations. Patient dropout is an important factor in longitudinal studies and must be accounted for when deriving sample size estimations. Other factors that have been explored in studies include within-patient correlations and the number of repeated measures (Liu and Liang 1997; Zhang and Ahn 2011).

Chapter 3

LMM and GEE

This chapter pertains to models containing continuous response variables only.

3.1 Linear Model

In order to more easily understand LMM and GEE models, an explanation of a simple linear model is provided. Assume that the response y_{ij} from patient $i = 1, ..., m$ at time $j = 1, ..., n_i$ is dependent on time t_j. This can be expressed as

$$y_{ij} = \beta_0 + \beta_1 t_j + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{3.1}$$

where the residual error ϵ_{ij} not explained by the model is normally distributed with zero mean and variance σ^2. It is assumed that all ϵ_{ij} are mutually independent. When considering longitudinal measurements performed on a group of patients it is usual that observations for patient i will be positively correlated. A model that doesn't account for this within-patient correlation can result in badly biased standard errors, although estimates for the model parameters will be unaffected (Agresti 2007, p276). Both LMM and marginal models address within-patient response correlation using two separate methods.

3.2 Linear Mixed Model (LMM)

For simplicity, LMM is described in terms of a single time explanatory covariate (and the intercept term).

3.2.1 Random intercept model

Consider a patient population that is monitored longitudinally, whereby the response variable is continuous and there is a response trend. The variability in the response outcome can be split into between-patient and within-patient variability. A key feature of LMMs is that the two types of variability are measured separately. LMMs are also known as conditional models. Where the trend is constant between patients, but the responses between patients differ, the LMM *random intercept model*, can take the form

$$\begin{aligned} y_{ij} =&(\beta_0 + b_{0i}) + \beta_1 t_j + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2) \\ & b_{0i} \sim N(0, \sigma_0^2) \end{aligned} \tag{3.2}$$

The response y_{ij} for patient $i = 1, ..., m$ at time $j = 1, ..., n_i$ is dependent on the mean (fixed) intercept β_0 and (fixed) slope (β_1) of the model, as well as the patient-specific contribution to the intercept a_{0i} and the residual error ϵ_{ij}. The response y_{ij} is therefore conditional on patient i specific intercept, $\beta_0 + b_{0i}$.

In equation 3.2, the residuals (ϵ_{ij}) are assumed to be mutually independent and arise from a normal distribution with zero mean and variance σ^2. Residuals may however be temporally correlated, although this scenario is not considered further in this report. Equation 3.2 also assumes that all patient-specific intercepts ($\beta_0 + b_{0i}$) subtracted from the population mean intercept (β_0) in the sampled population represent random effects that arise from a larger population of patients. The random intercept effect is assumed to be normally distributed with a zero mean and variance σ_0^2. The b_{0i} are assumed mutually independent. From equation 3.2 it can be seen that combining the between- and within-patient variance components result in the total model variance

$$Var(y_{ij}) = \sigma_0^2 + \sigma^2 \tag{3.3}$$

Because two measurements for patient i at times j and k are not independent the resulting co-variance between them is defined as

$$Cov(y_{ij}, y_{ik}) = \sigma_0^2$$

The intraclass correlation (ICC) between two measurements for patient i at times j and k in the random intercept model is therefore

$$ICC = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} \tag{3.4}$$

Because all the variation in an LMM is specified in terms of probability distributions, the joint probability distribution is therefore known. As a result, parameter estimation for LMMs can be performed using maximum likelihood (ML) or restricted maximum likelihood (REML).

3.2.2 Random slope model

In equation 3.2 only variation in the intercept term between patients was addressed. However, the trend (slope) between patients may also vary, and an LMM needs to account for this. Equation 3.5 incorporates this extra variation

$$\begin{aligned} y_{ij} =& (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2) \\ & (b_{0i}, b_{1i})^t \sim N(0, \Phi) \end{aligned} \tag{3.5}$$

$$\mathbf{\Phi} = \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} \tag{3.6}$$

where b_{1i} represents the random component of the slope for patient i and $(\beta_1 + b_{1i})$ represents the subject-specific random slope. It is assumed that the slope and intercept random effects have a bivariate normal distribution with covariance matrix Φ. Correlation ρ between intercept and slope terms may exist, but mutual independence between subject-specific random effects and residuals is assumed. The variance of the response for equation 3.5 is given by

$$Var(y_{ij}) = \sigma_0^2 + 2\rho\sigma_0\sigma_1 + \sigma_1^2 t_j^2 + \sigma^2 \tag{3.7}$$

and the co-variance given by

$$Cov(y_{ij}, y_{ik}) = \sigma_0^2 + \rho\sigma_0\sigma_1(t_j + t_k) + \sigma_1^2 t_j t_k, \tag{3.8}$$

indicating that constant between-patient covariance and hence correlation may not exist when utilising multiple random effects (Tango 2017, p69).

3.3 LMM matrix notation

In order to incorporate covariates such as the treatment group into LMM models, matrix notation is utilised. An LMM model can be described as

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{u} + \boldsymbol{\epsilon_i}$$

where $\boldsymbol{y_i}$ is the response vector for patient i, $\boldsymbol{X_i}$ is the the fixed effects covariate matrix for patient i, $\boldsymbol{\beta}$ is the vector of coefficients for the fixed effects, $\boldsymbol{Z_i}$ is the random effects covariate matrix for patient i, $\boldsymbol{u}$ is a vector of coefficients for the random effects and $\boldsymbol{\epsilon_i}$ is a vector of residual errors for patient i. This notation can be modified further to incorporate all patients $i = 1, ..., m$ by the following equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{3.9}$$

For simplicity it is assumed that the variances and covariances of random effects and residuals errors in the two treatment groups are equal.

3.4 Sample correlation

To appreciate the construction of a marginal model an understanding of the measurement of correlation between patient observation is required. The sampled correlation r_{jk} between within-patient responses for patients $i = 1, ..., m$ at time t_j and t_k can be estimated calculated utilising equation 3.10.

$$r_{jk} = \frac{\sum_{i=1}^{m}(y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{m}(y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{m}(y_{ik} - \bar{y}_k)^2}} \tag{3.10}$$

The terms y_j and y_k denote the mean response at time j and k respectfully.

3.5 Marginal GEE models

The term *marginal* arises from the fact that the mean response of a marginal model is dependent only on the covariates of interest. In contrast, for an LMM, not only is the response variable dependent on the covariates of interest but also on the random effects generated by the sampled population of patients (Fitzmaurice, Laird, and Ware 20012, p342).

For continuous data, equation 3.1 can be expressed as the marginal model for all patients $i = 1, .., m$ at time $j = 1, ..., n_i$ as

$$E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{3.11}$$

where $\boldsymbol{y}_i$ represents the vector of responses for patient $i = 1, ..., m$ at time $j = 1, ..., n_i$, $\boldsymbol{X}_i$ represents the covariate matrix for patient i at time j and $\boldsymbol{\beta}$ represents the matrix of covariate coefficients. The covariance matrix of the marginal model $Cov(\boldsymbol{y}_i)$ for patient i that incorporates the within-patient *working correlation matrix* $\boldsymbol{R}_i$ (section 3.5.1) is expressed by the formula

$$Cov(\boldsymbol{y}_i) = \sigma^2 \boldsymbol{R}_i \tag{3.12}$$

where

$$\boldsymbol{R}_i = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \cdots & \rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \cdots & 1 \end{pmatrix}, \tag{3.13}$$

ρ_{jk} is the correlation based on the pre-defined structure at time j and k, given $j, k = 1, ..., n_i, j \neq k$ and σ^2 represents the residual variance. While it is possible that the residual variance σ^2 of the model is dependent on the time j of observation, for simplicity it is assumed that σ^2 remains constant for time $j = 1, ..., n_i$.

3.5.1 Correlation structure

Marginal models incorporating repeated measures can be analysed utilising generalised estimating equations (GEE). GEEs represent a multivariate generalisation of the quasi-likelihood (section 3.5.2) for longitudinal data (Agresti 2007, p280). Having defined the relationship for $E(\boldsymbol{y}_i)$ and $Cov(\boldsymbol{y}_i)$, and assuming that σ^2 is normally distributed, the correlation structure $\boldsymbol{R}_i$ for patient i needs to be defined. This structure is known as the *working correlation matrix*. All diagonal elements of the *working correlation matrix* are assigned the value 1, and values of the off-diagonal components ρ_{jk} are dependent on the type of structure chosen. Four common *working correlation matrix* structures will be described.

The first correlation structure is termed *exchangeable* or *compound symmetry*. All components within the correlation matrix are assigned the same value, ie $\rho_{12} = \rho_{21} = \rho_{jk} = \rho$. Over an extended period of time, where correlation between time points t_j and t_k may decrease with time, another structure termed *auto regressive (AR)* (of order 1) may be more appropriate. Each component of the correlation matrix is given the value $\rho^{|j-k|}$. As the time between observations increase, the value of $\rho^{|j-k|}$ approaches zero. A third type of correlation structure is termed *independent*, whereby all correlation components have a value zero. This structure represents within-patient observations that are independent. A final correlation structure is known as *unstructured* and represents correlations observed from the data ie $\rho_{jk} = r_{jk}$ (section 3.4). While it may seem appealing

to utilise this structure in a marginal model, the disadvantage is that more parameters need to be estimated, thereby reducing the degrees of freedom available for model prediction. For example, when three time points have been utilised in the model three parameters need to be estimated. For exchangeable or AR structures only one parameter must be estimated.

It is worth noting that the correlation parameter ρ using an *exchangeable* symmetry is equivalent to the ICC in the random intercept LMM. In this case collapsing the combined covariance structure of the random intercept (b_{0i}) and residual variance (ϵ_{ij}) (see equation 3.2) will result in a matrix that contains the *exchangeable* correlation structure.

3.5.2 Quasi-Likelihood

Maximum likelihood (ML) and restricted maximum likelihood (REML) estimation methods of the regression coefficients for LMM are based on the assumption that the joint probability of responses for patient i are known. Another technique, the quasi-likelihood function, relies on the assumption that only the marginal (univariate) distribution of patient i response y_{ij} at time j is known (Hedeker and Gibbons 2006, p134). For example, when considering two continuous responses of patient i at time j and k ie. y_{ij} and y_{ik}, it would be assumed that the probability of each response would follow a univariate normal distribution, rather than a (joint) bivariate normal distribution. The regression equation is modeled separately from the covariance, and the covariance is considered a nuisance parameter. The advantage of utilising a quasi-likelihood method for parameter estimation is that knowledge of the joint probability distribution is not required. GEE utilise the quasi-likelihood estimation approach.

3.5.3 Parameter Estimates

Solutions to GEE can be obtained analytically for continuous data. Utilising equation 3.11 the values of the coefficient matrix $\hat{\boldsymbol{\beta}}$ are estimated from the vector of explanatory variables $\boldsymbol{X}_i$, response variable $\boldsymbol{y_i}$ and the $n_i \times n_i$ working correlation matrix $\boldsymbol{R_i}$ for patient i. The formula for estimating $\hat{\boldsymbol{\beta}}$ is (Hedeker and Gibbons 2006, p137)

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{m} \boldsymbol{X}_i^t[\boldsymbol{R_i}]^{-1}\boldsymbol{X}_i\right]^{-1}\left[\sum_{i=1}^{m} \boldsymbol{X}_i^t[\boldsymbol{R_i}]^{-1}\boldsymbol{y_i}\right] \tag{3.14}$$

In large samples, $\hat{\boldsymbol{\beta}}$ is found to be a consistent estimator of $\boldsymbol{\beta}$, independent of the working correlation matrix selected (Fitzmaurice, Laird, and Ware 20012, p357). In addition, in large samples the sampling distribution of $\hat{\boldsymbol{\beta}}$ is multivariate

normal with mean $\hat{\boldsymbol{\beta}}$ and $Var(\hat{\boldsymbol{\beta}})$ (Fitzmaurice, Laird, and Ware 20012, p358). The formula for $Var(\hat{\boldsymbol{\beta}})$ is

$$Var(\hat{\boldsymbol{\beta}}) = \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$$

where

$$\boldsymbol{B} = \left[\sum_{i=1}^{m} \boldsymbol{X}_i^t[\boldsymbol{R_i}]^{-1}\boldsymbol{X}_i\right]$$
$$\boldsymbol{M} = \left[\sum_{i=1}^{m} \boldsymbol{X}_i^t[\boldsymbol{R_i}]^{-1}\boldsymbol{X}_i(y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})(y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})^t[\boldsymbol{R_i}]^{-1}\boldsymbol{X}_i\right]$$

$Var(\hat{\boldsymbol{\beta}})$ is known as the robust or *sandwich estimator* (Hedeker and Gibbons 2006, p138). In many cases, even if the working correlation matrix is misspecified, valid standard errors of $\hat{\boldsymbol{\beta}}$ are obtained using sandwich estimator. The GEE estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator even with a misspecified working correlation matrix. For continuous response variables, the quasi-likelihood estimator is almost as precise and efficient as the maximum likelihood estimator (Fitzmaurice, Laird, and Ware 20012, p358).

3.6 Sample size formulas

Two simple sample size formulas have been developed for the LMM model with continuous outcomes (Liu and Liang 1997; Zhang and Ahn 2011; Diggle et al. 2002, p29–30; Fitzmaurice, Laird, and Ware 20012, chap. 20). The appropriate sample size formula is dependent upon the fitted regression lines between treatment groups. Assume that data can be modeled by the relationship

$$y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 g + \beta_3 (t_j \times g) \tag{3.15}$$

where t_j represents time $j = 1, ..., n_i$, $g = 0$ for the control group and $g = 1$ for the interventional group, and $(t_j \times g)$ represents the change in slopes (interaction) between the fitted regression lines of the two groups. If the fitted regression lines for the interventional and control group are not parallel, the interaction term coefficient β_3 is not null. Therefore the slope of the fitted regression line for the intervention group will differ from that of the control group. Assuming that (1) the correlation between any two time measurement within a patient is constant (ie exchangeable correlation structure), (2) the sample size is equal between groups, (3) responses for all patients are measured at the same time points, (4) the variance $Var(y_{ij})$ for both groups are equal and (5) $Var(y_{ij})$ of each model are normally distributed, the sample size formula is given as

$$m = \frac{2(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 Var(y_{ij})(1-\rho)}{s_t^2 d_s^2} \tag{3.16}$$

where m is the sample size, Z_p is the pth quantile of a standard Gaussian distribution, α is the type I error rate, β is the type II error rate and $power = 1-\beta$, $Var(y_{ij})$ is the variance within the model, ρ is the correlation between repeated measures, d_s is the CRD between the slopes of the two groups and $s_t^2 = \sum_j (t_j - \bar{t})$ for each time point $j = 1, ..., n_i$. $Var(y_{ij}) = \sigma_0^2 + \sigma^2$, where σ_0^2 is the between-patient random variation and σ^2 is the residual variance (see equation 3.3) and ρ is the ICC (see equation **??**). Equation 3.15 is valid for the LMM random intercept model. If a LMM is utilised and a random slope effect is found to exist in addition to a random intercept effect, then Fitzmaurice, Laird, and Ware (20012) (p586) states the variance component $Var(y_{ij})(1-\rho)/s_t^2$ of equation 3.16 is replaced by the term $\sigma^2/s_t^2 + \sigma_1^2$, where σ_1^2 (from equation 3.5) represents the random slope variance.

If the coefficient for the interaction term in equation 3.15 is zero, ie $\beta_3 = 0$, then the average response between groups may wished to be investigated. In this case the fitted regression lines for the two groups would be parallel and the *time averaged difference* (TAD) between groups can be estimated. Applying the same assumptions from equation 3.16 the sample size for a group m is

$$m = \frac{2(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 Var(y_{ij})(1+(n-1)\rho)}{n d_m^2} \tag{3.17}$$

where d_m is the CRD of mean response between the interventional and control groups.

The structure of the covariance matrix for the GEE exchangeable model is equivalent to marginalising the random effects and residual variance matrices ($\boldsymbol{Zu}+\boldsymbol{\epsilon}$, see equation 3.9) of the random intercept model (equation 3.2). Therefore, when investigation TAD between treatment groups, equation 3.17 also applies to the GEE exchangeable model. However, parameter estimations in GEE models are based on large sample sizes. Thus the approximate sample size calculation for TAD using the GEE exchangeable model (equation 3.17) is based on a large sample normal approximation.

3.7 Comparison between LMM and GEE

GEE are utilised to analyse marginal models, which model the average population response. LMM in contrast are designed to model patient-specific effects. For the purpose of generating patient simulated data, LMM is the method of choice.

Both GEE and LMM accommodate missing data in their models. GEE models can accommodate data that is based on the strong, and often unrealistic missing completely at random (MCAR) assumption (Fitzmaurice, Laird, and Ware 20012, p358). In contrast, LMM are able to accommodate both MCAR and the weaker missing at random (MAR) assumption (Fitzmaurice, Laird, and Ware 20012, p497).

An assumption of LMM for continuous data is that realisations of the outcome response variable y_i for patient i follow a multivariate gaussian distribution across all time points j. Marginal models on the other hand only require that realisations of the outcome variable y_{ij} follow a univariate normal distribution specific to time point j. It is easier to predict the univariate probability distribution of the outcome variable at a single time point than the joint probability distribution for all time points. Therefore GEE models are more flexible than LMM to misspecifications of the probability distributions for the response variable. However, GEE models are based on large sample size normal distribution approximations and are likely to lead to increased type I error rates with small sample sizes (Morrel et al. 2009).

3.8 Review of repeated measures cRCT

Repeated measures cluster RCTs (cRCT) are a special type of repeated measures design that will be briefly described.

3.8.1 Definition

For non cRCTs, the unit of randomisation (cluster) is the individual. For cRCTs, the unit of randomisation is a group of individuals. Consider an RCT that compares two interventions by health visitors on mothers to reduce post-natal depression. Post-natal depression scores of mothers are measures at multiple time points. Each health visitor investigates a cohort of mothers. An intervention is randomly allocated to a health visitor assigned to a cohort of patients. This is three level hierarchical cRCT design. At the highest level (Level 3), the cRCT aims to detect differences in the post-natal scores of cohorts measured by the two groups of health visitors. For each cohort managed by a health visitor (Level 2) post-natal scores are analysed at the individual level. For each individual (Level 1), post-natal scores acquired at different time points are analysed.

3.8.2 Purpose

Clustered RCTs are designed for reasons that have been described by Hayes and Moulton (2016) (p5). Firstly, the nature of the intervention may require it to be applied to communities or other groups of individuals. It may be for convenience reasons. Secondly, without a cRCT design, contamination may occur. Contamination occurs when, for example, patients allocated to a health

professional receive the same intervention rather than the intervention that was randomised to the patient. This "contamination" will result in a reduction in outcome differences between interventional groups, leading to artificially reduced CRDs. Thirdly, population and community effects of an intervention may be primarily of interest.

3.8.3 Analysis

Repeated measures cRCTs, containing three organisational levels, lend themselves to be analysed using multilevel modeling. Both LMM (Moerbeek and Teerenstra 2015, chap. 9; Magnusson, Andersson, and Carlbring 2018) or GEE models (Guthrie et al. 2012) can be utilised for this purpose, with the advantages of each method having been previously discussed in this chapter. Alternatively, summary measures such as the post-randomisation mean can be used as the unit of measurement for the within-patient measurement level (Level 1), and a two-level hierarchical model employed. Using the example in section 3.8.1, this would mean using the summary measure for each patient as level 1 of the hierarchical model, and the cohort assigned to a health visitor as level 2 (S.Walters, personal communication).

3.8.4 Sample size estimation

With this extra level of complexity, sample size formulas become more complicated. Examples of three level hierarchical sample size formulas are discussed by Moerbeek and Teerenstra (2015) (chapter 9). Magnusson, Andersson, and Carlbring (2018) explored the effects of factors such as unbalanced designs and missing data on their analytical formulas. After verification with power analysis simulation studies, they concluded their overall sample size calculations were robust. However, power analysis remains the gold standard for sample size estimation and can be used to optimise sample size for a particular study. Sensitivity analysis using power analysis is an important tool that has been utilised to consider the effect of parameters on study design (Lane and Hennes 2019).

Chapter 4

Method development for power analysis

4.1 Introduction

Within this dissertation, power analysis was performed using simulation. The aims of this chapter are the following:

1. To describe details of the study by Thomas et al. (2006)
2. To determine the parameters of a model that fit the data of Thomas et al. (2006)
3. To describe the method of power analysis based on the parameters acquired from the model in (2)
4. To list factors that will be investigated in this study that affect power

4.2 Study data

4.2.1 Outcome measure and study design

Outcomes measured in the study by Thomas et al. (2006) were based on the quality of life (QoL) instrument SF-36 bodily pain index. The index ranged in single integer values from 2 to 11. This index was scaled from 0 to 100 (no pain) and contained 10 values (0, 11.1, 22.2,. . . ,100). Body pain scores were obtained from patients at baseline, 3 months, 12 months and 24 months.

The study was a parallel group design RCT and compared the effect of acupuncture and usual care on the pain outcomes (measured by the QoL index) of patients. Patients received traditional acupuncture for 10 sessions between baseline and 3 months from one of six registered acupuncturists. Those in the usual-care group received standard care over this period. Only standard care was administered after 3 months, and only if requested by the patient.

4.2.2 Sample size

A total of 159 patients, who were randomised to the "acupuncture" group and 80 patients, who were randomised to the "usual care" group had baseline measurements recorded. Those patients who did not have results recorded at 3, 12 and 24 months were eliminated from the analysis. As a result, 153 patients in the acupuncture group and 76 patients from the usual care group remained.

The number of patient outcomes recorded for both groups at different time points are shown in Tables 4.1 and 4.2

Table 4.1: Outcomes recorded at the sampled time points for acupuncture group

baseline	3months	12months	24months
153	146	147	123

Table 4.2: Outcomes recorded at the sampled time points for usual care group

baseline	3months	12months	24months
76	71	68	59

Within the acupuncture group 7 missing values were recorded at 3 months, 6 missing values at 12 months and 30 missing values at 24 months. Within the usual care group 5 missing values were recorded at 3 months, 8 missing values at 12 months and 17 missing values at 24 months. Missing value rates were within expectations (10-15%) at 3 and 12 months . The dropout rates were similar between groups at 24 months.

4.2.3 Exploratory Data Analysis

4.2.3.1 Mean QoL scores

A plot of the mean outcomes for each treatment group at the four time points is displayed in Figure 4.1

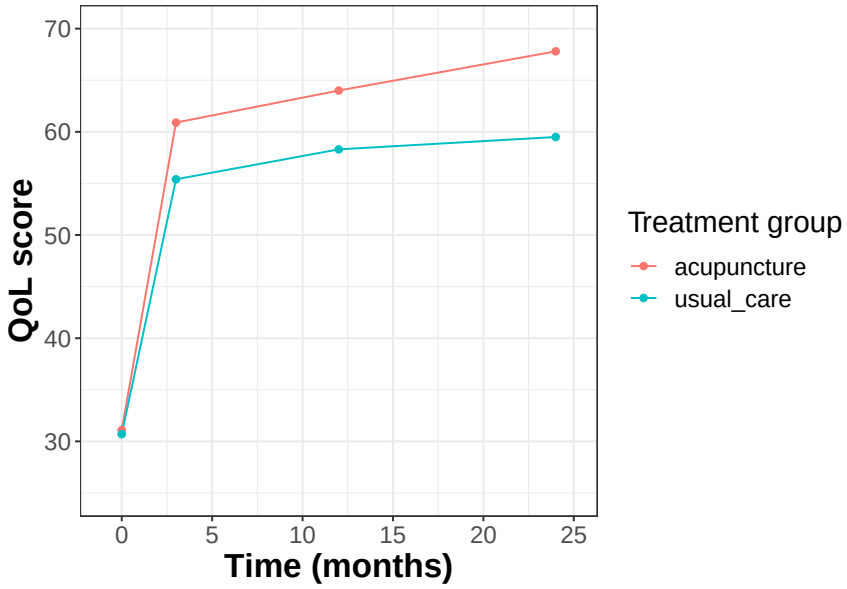

Figure 4.1: Mean QoL score between the two treatment groups. Comparisons are performed excluding baseline values. The mean QoL score for the acupuncture group is higher than for the usual care group. A linear relationship appears to exist between time and mean QoL score. The differences between the two groups appear constant

The mean QoL score at baseline for the acupuncture group was 31.1 compared to 30.7 for the usual care group. No significant statistical difference in mean QoL between these two groups was found using a t-test (Thomas et al. 2006). This is not surprising as patients were randomly selected for both treatment groups.

After the interventional treatment was completed, the mean QoL of both groups increased substantially, suggesting that both treatments were effective compared to no treatment. The mean QoL value for the acupuncture group at each time point was greater than for the usual care group. The mean QoL score for the acupuncture group appeared to increase in a linear manner with time, while for the usual care group the the mean QoL appeared to flatten out. A possible interaction between time and group may have existed according to figure 4.1.

4.2.3.2 Distribution of QoL scores

The QoL outcome response utilised in this study is discrete. In this form, it is technically not valid to utilise models that assume continuous response data. Walters (2009) (p69) argues that if there are more than around 7 discrete values on the outcome scale, and that the data is relatively symmetric with a proportionally low number of values at the upper/lower bounds, then the data can modeled as continuous. Figure 4.2 demonstrates the distribution of QOL values at the 3, 12 and 24 month time points.

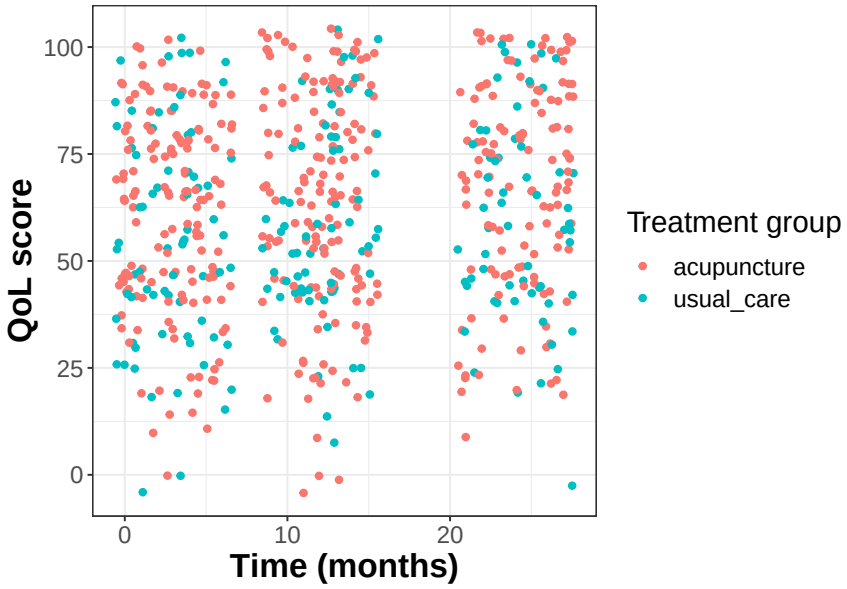

Figure 4.2: Distribution of QOL scores at baseline, 3, 12 and 24 months. Values at each time point have been jittered for ease of visualisation.

There appears to be a low proportion of low QoL values for all time points. The spread of data otherwise appears uniform across QoL values. As such the response variable was considered to have come from a continuous distribution, as suggested by Walters (2009) (p69).

4.3 The statistical model

4.3.1 Choice of Model

An LMM was chosen to model the data of Thomas et al. (2006) because an LMM can account for individual random effects. Simulations of new datasets containing individual responses at various time points are easily generated using an LMM. The package *lme4* (version 1.1-23) in R (version 3.6.2) was utilised for this purpose.

4.3.2 Explanatory variables

In the study by Thomas et al. (2006) a range of covariates were investigated. These included *age*, *sex*, *acupuncturist*, *duration in current episode of lower back pain (LBP) in weeks*, *expectations of back pain in 6 months* , *SF-36 Physical Functioning* , *reported pain in legs* and *reported pain in back*. Only the baseline measurement was recommended for inclusion as a covariate in the LMM (S. Walters, personal communication).

4.3.3 Variable re-scaling

In order to improve numerical precision for model calculations, the outcome response (QoL score) and the time explanatory variables were adjusted to a similar scale. The outcome variable divided by 100 resulted in an output range between 0 and 1. The time variable was expressed in years, with time zero (t_0) = 3 months, $t_{0.75}$ = 12 months and $t_{1.75}$ = 24 months.

4.3.4 Inclusion of explanatory effects

4.3.4.1 Analysis of random effects

The two random effects investigated in the LMM were time and the intercept. The inclusion of random effects in the model was determined by comparing the nested LMMs (with and without slope random effect) utilising the full model of fixed effects (see appendix A for R code). Each model was evaluated using *restricted maximum likelihood* (REML), as this provided an unbiased estimate of the variance components. A *generalised likelihood ratio test* (GLRT) was performed and the test statistic was evaluated using a χ^2 test (see appendix A for the R code). It is noted that a p-value obtained from this test can be conservative (Pinheiro and Bates 2000, p84). One method of correcting for the conservatism of the GLRT is to divide the p-value by 2 (Kain, Bolker, and McCoy 2015). The level of statistical significance was taken at the $p = 0.05$ level. It was found that the inclusion of the slope and slope-intercept covariate terms were not significant ($\chi^2_2 = 1.04$, p=0.37), even if the p-value was divided by 2 (p=0.18).

4.3.4.2 Analysis of fixed effects

From figure 4.1 it was suggested that the variables *baseline*, *time*, *group* and the interaction term $time \times group$ should be included in the full model of fixed effects. When comparing nested fixed effects models using the GLRT, the probability distribution of the test statistic is only approximated by the the χ^2 distribution. This can result in anti-conservative p-values (Pinheiro and Bates 2000, p86). In order to discover a more accurate probability distribution for the test statistic, parametric bootstrapping was performed. In parametric bootstrapping, data is simulated based on the reduced model and a test statistic calculated. This is repeated many times in order to develop a distribution of the test statistic based on the null hypothesis of the reduced model. It is noted that parametric bootstrapping is computationally expensive.

To perform hypothesis testing of nested fixed effects models, it was necessary that each model was estimated using *maximum likelihood* (ML) rather than REML (Pinheiro and Bates 2000, p87). The significance of the $time \times group$ interaction term was assessed by comparing the following two nested models,

$$response = baseline + time + treatment + time \times treatment + (1|patient)$$
$$response = baseline + time + treatment + (1|patient)$$

where $(1|patient)$ represents the random intercept effect (see appendix A for R code). The p-values found for the difference between these two models utilising parametric bootstrapping and the χ^2_1 distribution were 0.35 and 0.32 respectively. As the interaction term was not found to be significant it was dropped from the model.

Following a similar argument, the time term was found to be significant (p = 0.015 using parametric bootstrapping and p = 0.003 using the χ^2_1 distribution) and therefore kept in the model. The group term was also found to be significant (p=0.035 using parametric bootstrapping and p=0.026 using the χ^2_1 distribution) and thus retained in the model. Finally, the baseline term was found to be significant and and also retained in the model (p=0.00 using both parametric bootstrapping and the χ^2_1 distribution). The LMM that best described the observed data was

$$response = (0.4417 + b_0) + 0.3614 \times baseline + 0.0287 \times time + 0.0610 \times treatment + \epsilon \tag{4.1}$$

where $b_0 \sim N(0, 0.1625)$ and $\epsilon \sim N(0, 0.1661)$. The estimate for the increase in the scaled QoL index for the acupuncture group compared to the usual care group was 0.0610, translating to a QoL increase of 6.1 units.

4.4 Model checking

Checks on the assumptions of the LMM expressed by equation 4.1 are described in appendix B.

4.5 Simulations

4.5.1 Power analysis

The R code for creating a simulated dataset from the model parameters obtained in equation 4.1 is contained in appendix C. A GLRT was performed between a full model (LMM or GEE) containing the baseline, time and treatment variables, and a reduced model that contained only the baseline and time variables. It was impractical to obtain a parametric bootstrap distribution for each simulated dataset to test the significance of the GLRT statistic, due to computational speed issues. Instead a χ^2_1 distribution was utilised to determine the significance of the

test statistic. As has been discussed previously, utilising a χ^2 distribution for this task may result in an excess number of recorded significant results (ie. the power). For each set of values of factors investigated, 1000 simulations were performed. This number of simulations has been used in equivalent studies (Bahçecitapar 2018; Brandmaier et al. 2018).

Hypothesis testing of LMMs, as well as GEE models based on the exchangeable, AR, independent and unstructured covariance structures were performed on the simulated data. For the most part, sample sizes varied from 25 to 500, incremented by 25. The CRD for the full model was simulated at values 0,2,3,5 and 8 QoL units. Sample size was the same for each treatment group. For convenience, the default between-subject variation parameter was set at $b_1 \sim N(0, 0.1661)$ for LMM, resulting in the ICC equal to 0.5. All simulations were performed utilising a unit variance of $\epsilon \sim N(0, 0.1661)$.

4.5.1.1 Type I error assessment

The nominal type I error (α) for all simulations was set at 0.05. The simulated type I error was determined by measuring the power of simulations at $CRD = 0$.

4.5.1.2 Factors

Simulations were performed for a multitude of factors on each model included the following :

- ICC values of 0, 0.2, 0.5 and 0.8 utilising 3 post-baseline time-points (3,12 and 24 months)
- small treatment sample sizes from 5 to 50 in increments of 5
- 2-6 repeated measurements post-baseline over a fixed 21 month duration. sampled time points at repeated measurements were as follows:
  - 2: 3 and 24 months
  - 3: 3, 12 and 24 months
  - 4: 3, 10, 17 and 24 months
  - 5: 3, 8, 13, 18 and 24 months
  - 6: 3, 7, 11, 15, 19 and 24 months
- Utilising 3 post-baseline time points (3,12 and 24 months) missing values were assigned to a 10% random sample at 12 months and 20% random sample at 24 months.

4.5.1.3 Theoretical power curve

Where applied in the analysis, the theoretical power curve was generated from the formula expressed in equation 3.17. The relevant R code is provided in appendix C.

Chapter 5

Results and Discussion

5.1 Introduction

An aim of this investigation is to determine the effect of altering various parameters on the study power. The results of these investigations are described and discussed below in the following order:

1. the effect on Type I error by sample size, model selection, correlation and missing data pattern
2. the effect of model selection at an ICC value of 0.5
3. the effect of small sample size
4. the effect of ICC
5. the effect of GEE model covariance specification on power using an ICC of 0 to accurately reflect the GEE independent model
6. the effect of the number of sampled time points within a fixed period
7. the effect of missing data

5.2 Type I error

The type I error was visually investigated for all simulations. Figure 5.1 demonstrates the empirical type I error for simulations using between 2-6 sampled time points and an $ICC = 0.5$. The nominal type I error $\alpha = 0.05$ is indicated by the dashed line. No simulation was possible with the unstructured GEE model utilising 2 sampling points. For all models the type I error fluctuated predominantly between 0.03 and 0.07. At sample sizes above 50, the type I error typically appeared to stabilise and fluctuate around the nominal value. For sample sizes below 50, several models displayed an increase in type I error. The unstructured GEE model for all time point combinations displayed large type I errors for sample size of 25, especially using 5 and 6 time points. The independent GEE model with 6 time points, the LMM with 5 time points and the AR GEE model with 2 time points also had type I errors greater than 0.07

for a sample size of 25. There didn't appear to be a relationship between the initial type I error and number of time points sampled.

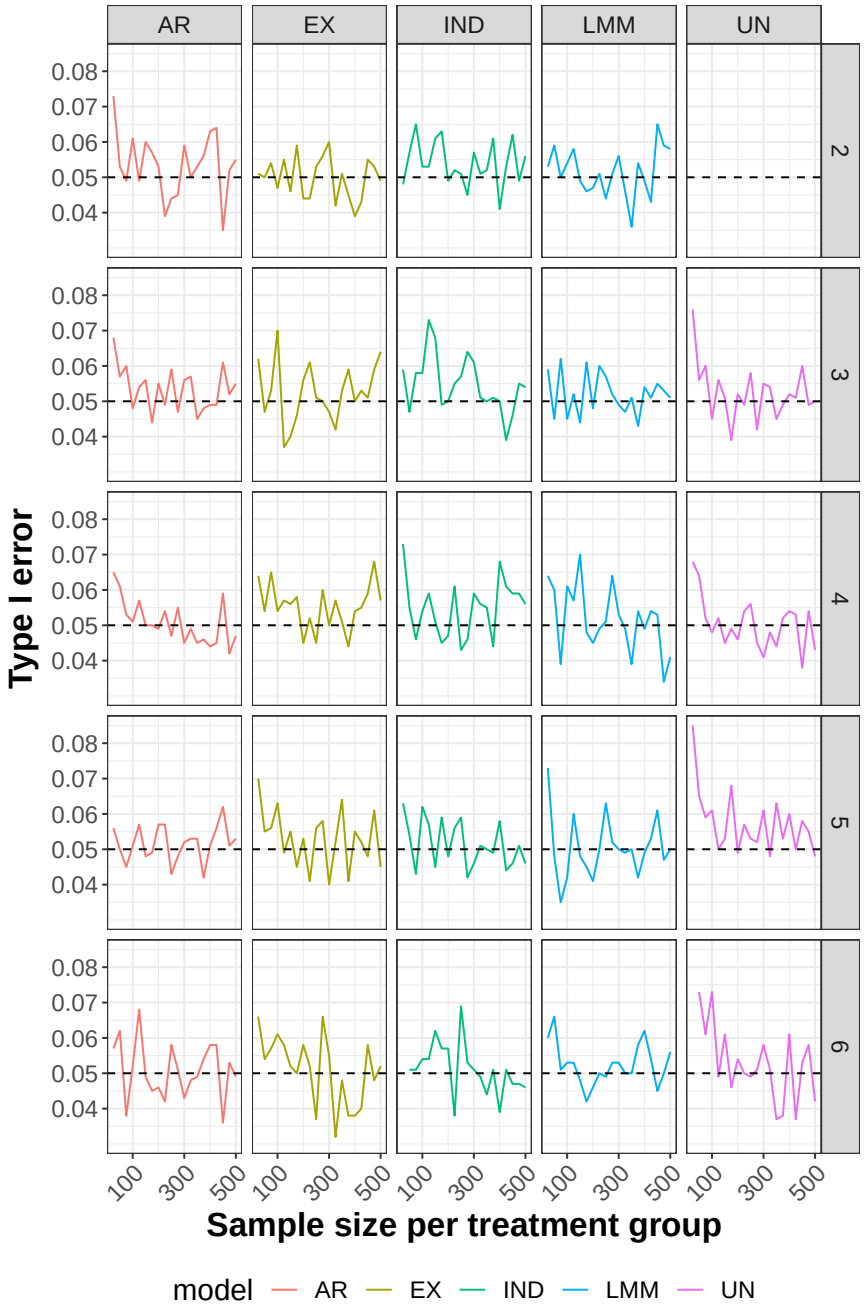

Figure 5.1: Type I error simulations conducted using 2-6 time points using LMM and GEE models. ICC value for these simulations was 0.5. It was not possible to conduct the simulation for the unstructured model using 2 time points. The dashed line represents the nominal 0.05 type I error. AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured

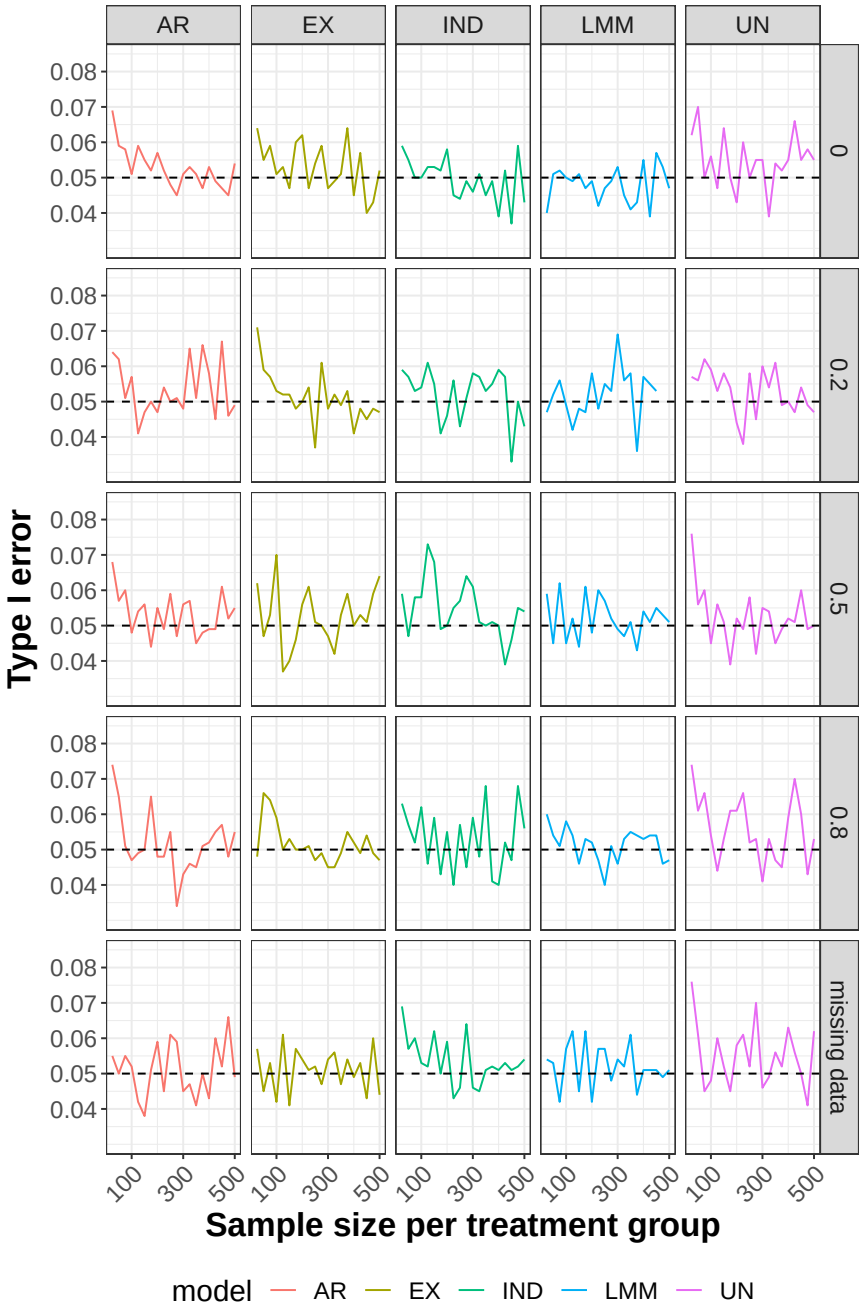

Figure 5.2: Type I error simulations conducted using ICC values 0, 0.2, 0.5 and 0.8. Each simulation utilised 3 time points. Missing data simulations were conducted with an ICC value of 0.5. For this plot, 10 percent of points at 12 months and 20 percent of points at 24 months contained missing values. The dashed line represents the nominal 0.05 type I error. AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured

The type I errors for the models at various ICC values, including that for missing value simulation ($\rho = 0.5$), and using 3 time points are displayed in Figure

5.2. For the unstructured model, at ICC values 0.5 and 0.8 as well as including missing values ($ICC = 0.5$), the type I error is greater than 0.07 at sample size of 25. Other models where the type I error is greater than 0.07 at sample size 25 is the AR model ($ICC = 0.8$) and the exchangeable model ($ICC = 0.2$). There doesn't appear to be a relationship between ICC and raised initial type I error.

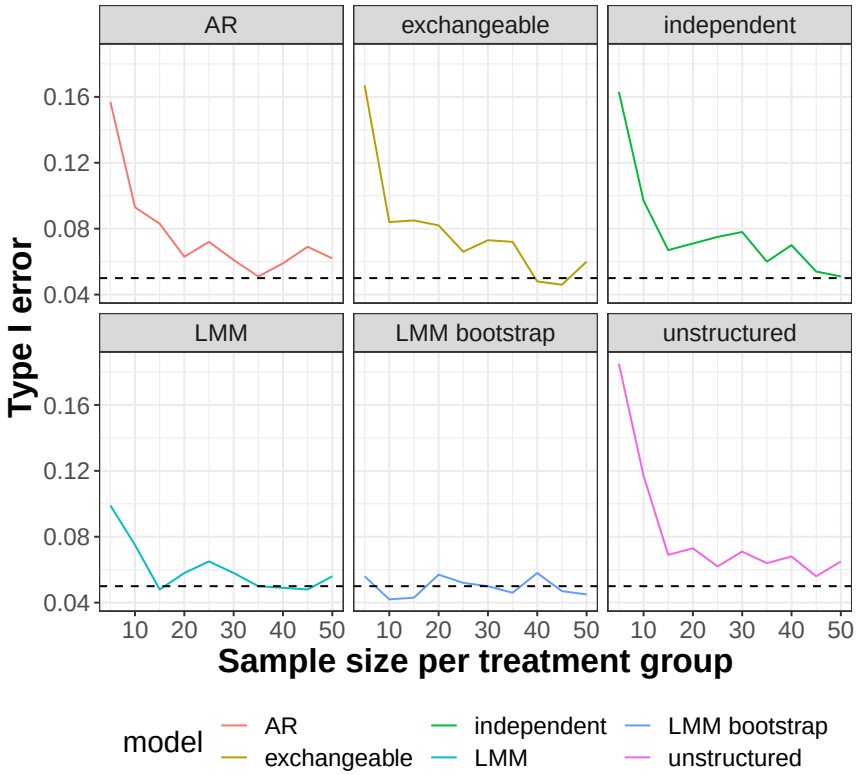

Figure 5.3: Type I errors for low sample sizes of all models acquired using a chi-squared test are displayed in conjunction with the type I error for the random intercept model acquired with parameteric bootstrapping (LMM bootstrap). The type I error for the LMM bootstrap model appeared stable and fluctuated around the the nominal value. The dashed line represents the nominal type I error of 0.05.

Examination of figure 5.1 and figure 5.2 suggests that there is a raised type I error at low sample sizes, especially for the unstructured model. Because of the elevated type I errors at small sample sizes for many of the scenarios seen in figure 5.1 and figure 5.2, simulations were performed using sample sizes ranging from 5 to 50.

The effect of small sample sizes on the type I error can be seen in figure 5.3. For the very small sample size of n=5, there is a large increase in type I error above the nominal 0.05 value. This effect is greater for the GEE models, especially for the unstructured model. The type I error decreases with increasing sample size and at a sample size of 50 the simulated error for all models appear to converge

to the nominal type I error.

In order to explain the large increase in type I error at small sample sizes for all models, the simulation using the LMM was repeated utilising parametric bootstrapping. This generated an empirical distribution of the GLRT test statistic for the null hypothesis. As expected the type I error utilising bootstrapping was stable for all sample size values. The simulated type I error using the LMM and bootstrapping was 0.050 ± 0.0059 (mean$\pm$sd). The associated 95% confidence interval was (0.038, 0.061). According to Wilks theorem, the GLRT statistic will be χ^2_1 distributed only as the sample size $\to \infty$. Especially at low sample sizes, the p-value generated from the hypothesis test for fixed effects terms between nested models using the χ^2_1 probability distribution will be smaller than the empirically derived value using bootstrapping (Pinheiro and Bates 2000, p86). As a result, more significant p-values will be generated using the χ^2_1 probability distribution. Thus a larger type I error will be generated compared with the use of parametric bootstrapping.

Type I errors have been documented in the literature for small sample sizes using continuous response data. Simulation studies using GEE models have resulted in Type I errors up to 0.26 for sample size 10 (Morel, Bokossa, and Neerchal 2003). LMM models are believed to outperform GEE models at low sample sizes (McNeish and Harring 2017) and the simulations performed in this report support this belief.

The structure of the working correlation matrix in GEE models is poorly estimated with very small sample sizes (Skene and Kenward 2010). Reducing the number of variance parameters in the matrix can reduce this problem. A poorly estimated variance matrix may be the reason why the type I error in GEE models is greater than in the LMM with small samples. The fact that the unstructured covariance matrix requires estimation of more covariance parameters than other GEE models may explain why the type I error is greatest at small sample sizes using the unstructured GEE model.

The impact of an inflated type I error is to unknowing increase the probability of rejecting the null hypothesis when the null hypothesis is true. Inflating the type I error will result in an artificially inflated power value, according to equation 3.17. Where an inflated type I error has occurred incorrect conclusions about the null hypothesis may be drawn.

5.3 Power curves at ICC = 0.5

Qualitative assessment was performed on simulations with an $ICC = 0.5$ utilising the three post-baseline treatments (3,12, 24 months). The power functions for all four CRD (2,3,5,8) are displayed in figure 5.4. Assuming that the level of power required for the study is 0.9, only simulations utilising a CRD of 5 and 8 QoL units achieved this. The power analysis for all five models appeared similar. As the value of CRD increased, the slope of the power curve increased. As the value

of the power curve approached 1, the power curve flattened to the horizontal. There appeared minimal variation between models when the power curve slope was steep.

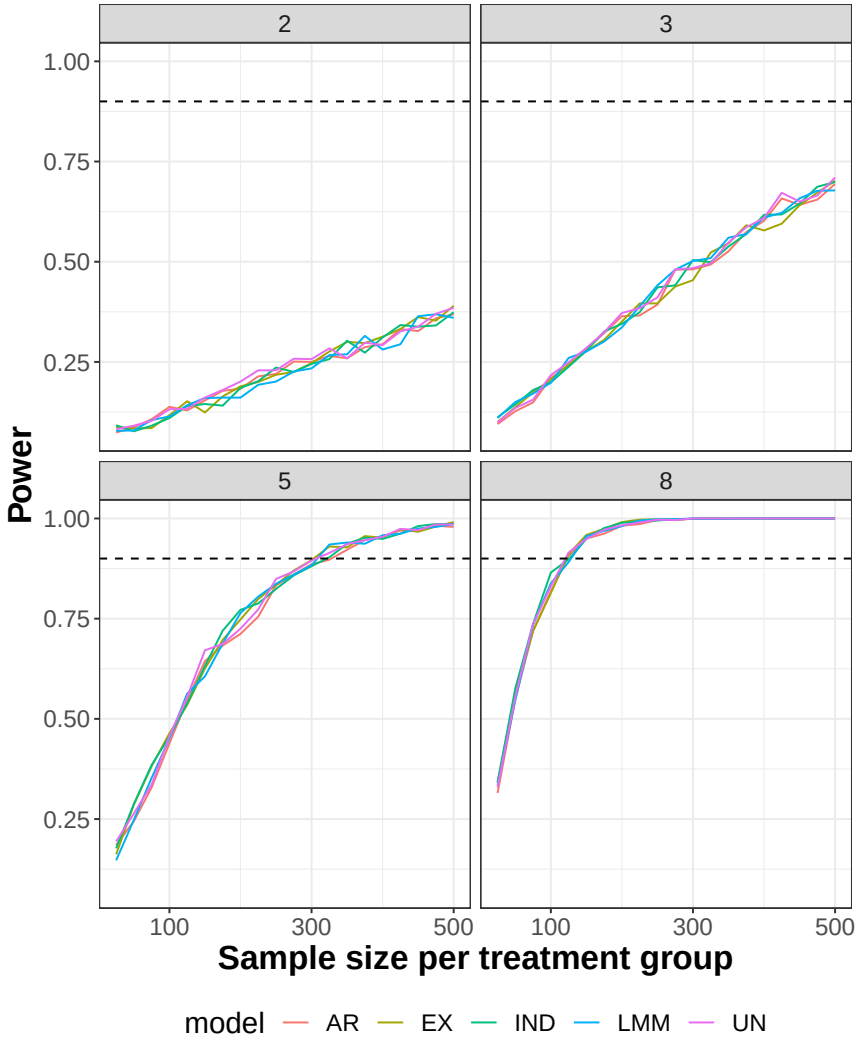

Figure 5.4: Power functions for the five models for the CRD of 2,3,5 and 8 QoL units. The dashed line represents a power level of 0.9. AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured

A comparison between the theoretically derived power curve (3.17) and simulated power curves for the CRD of 5 QoL units can be seen in figure 5.5. There does not appear to be any major discrepancies between the theoretical and simulated power curves. This observation suggests that the baseline covariate had little influence on the power calculation. This is unsurprising in an RCT because the patients are randomly allocated to groups and the mean responses of groups at baseline should not be different. This was inferred in the study by Thomas et al. (2006), where the acupuncture group had a mean (standard deviation) baseline

QoL score of 30.8 (16.2) compared to the usual care group mean baseline score of 30.4 (18.0).

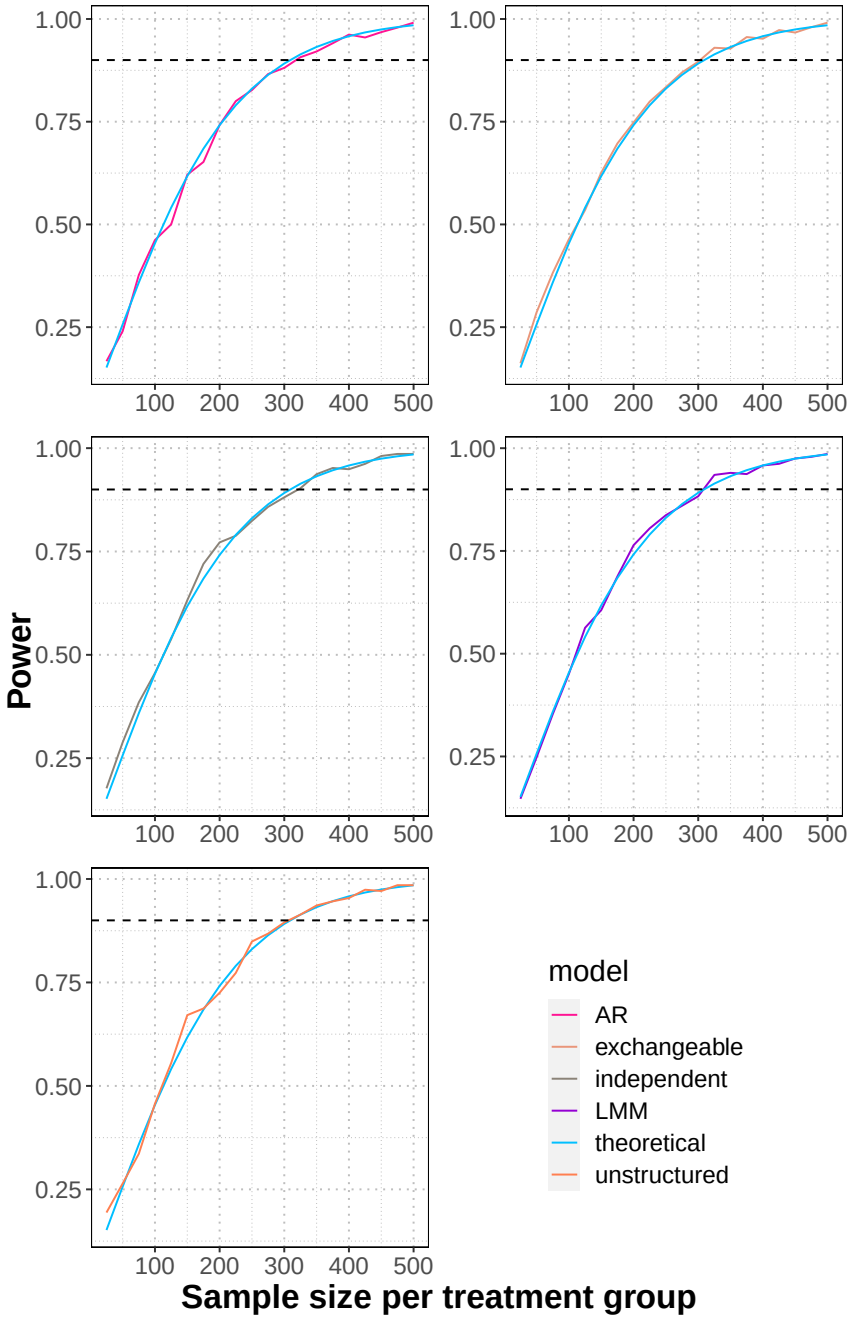

Figure 5.5: A comparison between the theoretical and simulated power curves with CRD of 5 QoL units and ICC of 0.5 for each model. From top left downwards the models are AR, independent, and unstructured, and from top right downwards the models are exchangeable and LMM. The dashed horizontal line represents the 0.9 power level.

5.4 Low sample size

Simulations for power analysis using all 5 models models at low sample sizes for each CRD is displayed in figure 5.6. Power curves at all CRD values appear as straight lines, except at very small sample sizes. From evidence seen in figure 5.3, relatively large power values in figure 5.6 were due to the large type I errors generated by models at sample sizes at or below 10. For sample sizes above 10 there is was visual evidence from figure 5.6 that a specific model outperformed another.

As the value of the CRD increased, so did the gradient of the power curves. At a CRD of 8 QoL units the power curves reached of power of 0.55. For this report it was assumed the desired power of a study was 0.9, well above the values seen in figure 5.6.

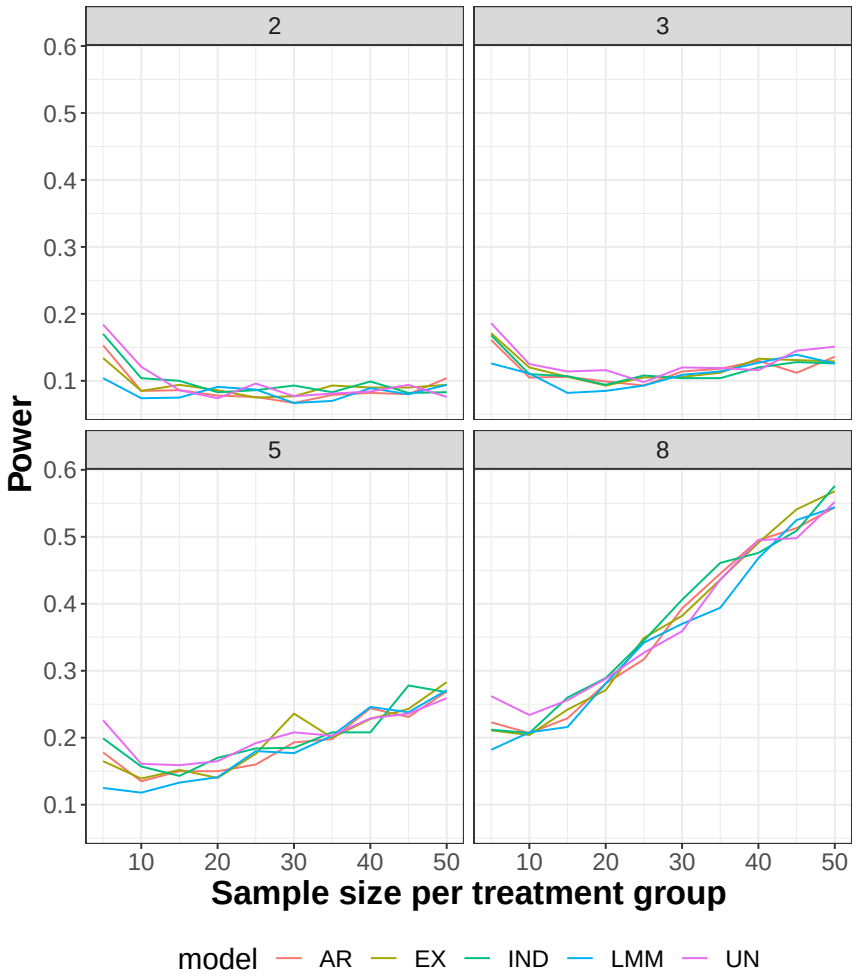

Figure 5.6: Power curves for LMMs using small sample sizes, CRDs 2,3,5 and 8 QoL units at an ICC of 0.5. AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured

5.5 Correlation

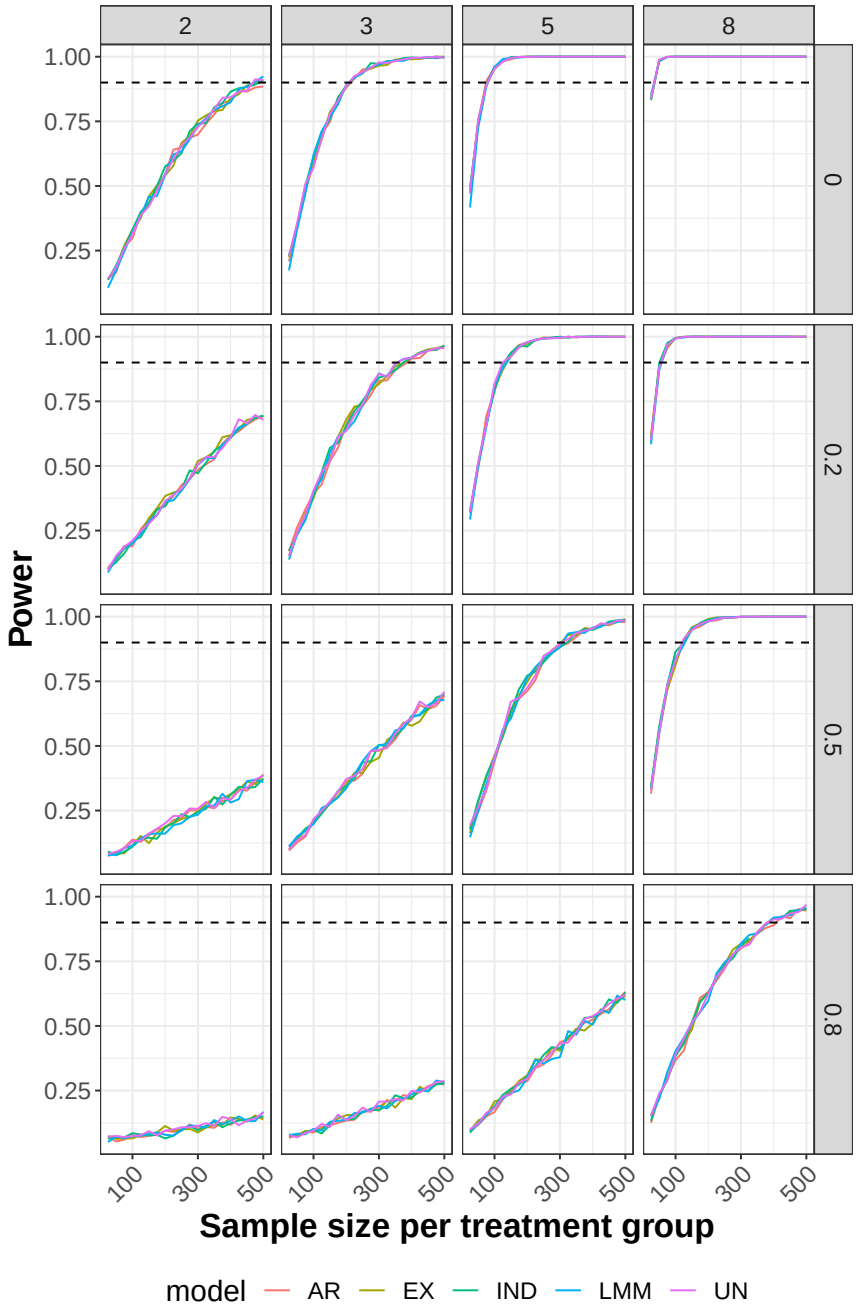

Figure 5.7: Power analysis using ICC values 0, 0.2, 0.5 and 0.8 (along vertical axis) and CRDs of 2, 3, 5 and 8 QoL units (along horizontal axis) for each model. Three time points were utilised for the simulations. Dashed line represents 0.9 power level. AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured

The effect of varying the ICC for the 5 models based on the three time points (3,12,24 months) can be seen in figure 5.7. There is no indication that one

Table 5.1: Power at values approximately 0.9, obtained from simulations using all models at a range of ICC, CRD and sample size values. (AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured)

			Models				
CRD	ICC	sample_size	LMM	EX	AR	IND	UN
2	0	475	0.90	0.90	0.88	0.89	0.91
3		225	0.92	0.92	0.92	0.92	0.92
	0.2	375	0.91	0.90	0.89	0.90	0.91
5	0	75	0.88	0.89	0.91	0.89	0.89
	0.2	125	0.88	0.87	0.90	0.89	0.90
	0.5	300	0.88	0.90	0.89	0.88	0.90
8	0	25	0.83	0.83	0.84	0.85	0.84
	0.2	50	0.87	0.88	0.88	0.90	0.88
	0.5	125	0.89	0.91	0.91	0.90	0.92
	0.8	375	0.89	0.89	0.88	0.89	0.90

particular model outperforms others at different values of ICC and CRD. It appears that the plots along the diagonal of figure 5.7 have a similar appearance. This is not surprising based on equation 3.17, where $\rho = ICC$. There is a complex inverse relationship between ICC and the d_m^2. Assuming that all other parameters in equation 3.17 are fixed, increasing ICC and d_m by specific quantities will result in little change of the power curve. For example, increasing CRD from 2 to 3 QoL units and ICC from 0 to 0.2 appear to result in minimal changes to the power curves (figure 5.7). These simulations indicate that increasing the ICC will result in a decreasing power when all other conditions remain constant.

Not all power curves in figure 5.7 crossed the 0.9 power threshold. Those that did for a given CRD and ICC are listed in Table 5.1. The range of different power values recorded between models at any combination of CRD and ICC was a maximum of 0.03. The difference observed between models at all combinations is not major, and likely to reflect sampling simulation error. Model systemic bias was not observed in table 5.1 as no model had power values consistently above or below others for the combination of CRD and ICC values.

5.6 Independent observations

The effect of correct specification of the covariance matrix in a GEE model on power was investigated utilising an $ICC = 0$. In this scenario, the independent working correlation correctly specifies the independently defined observations. Figure 5.8 displays the resulting power curves for CRDs of 2,3,5 and 8 QoL units. There appears little evidence that any GEE model or LMM is outperforming another. As the CRD increases, the variation between models is reduced. A

possible reason as to why the independent GEE doesn't outperform the other models is due to the robust sandwich estimator of the GEE model. Valid standard errors for the GEE model are obtained even if the correlation structure has been misspecified (Fitzmaurice, Laird, and Ware 20012, p358).

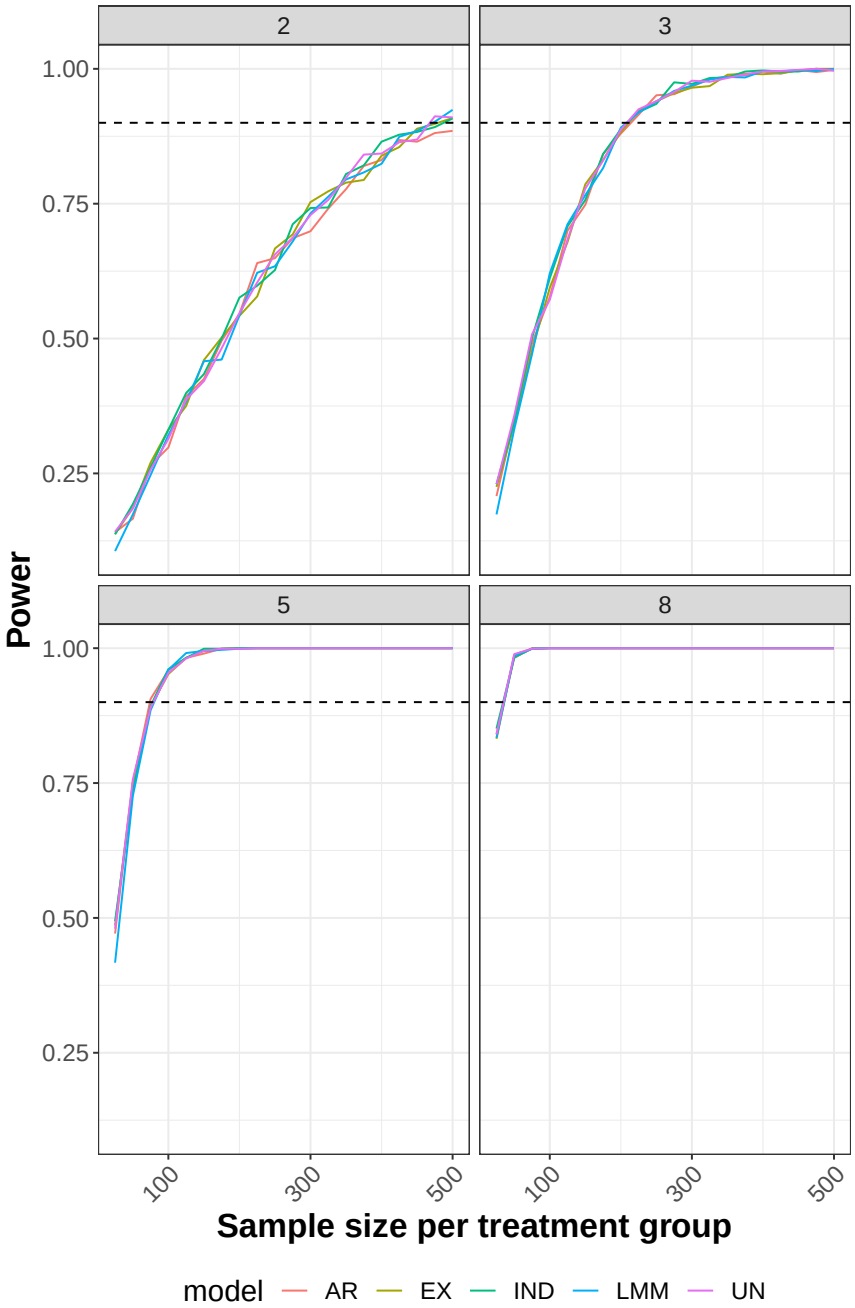

Figure 5.8: Power analysis using all models, and ICC = 0 and CRD values of 2, 3, 5 and 8 QoL units (each CRD represented by an individual plot). The dashed line represents a power level of 0.9. AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured

5.7 Number of sampled time points

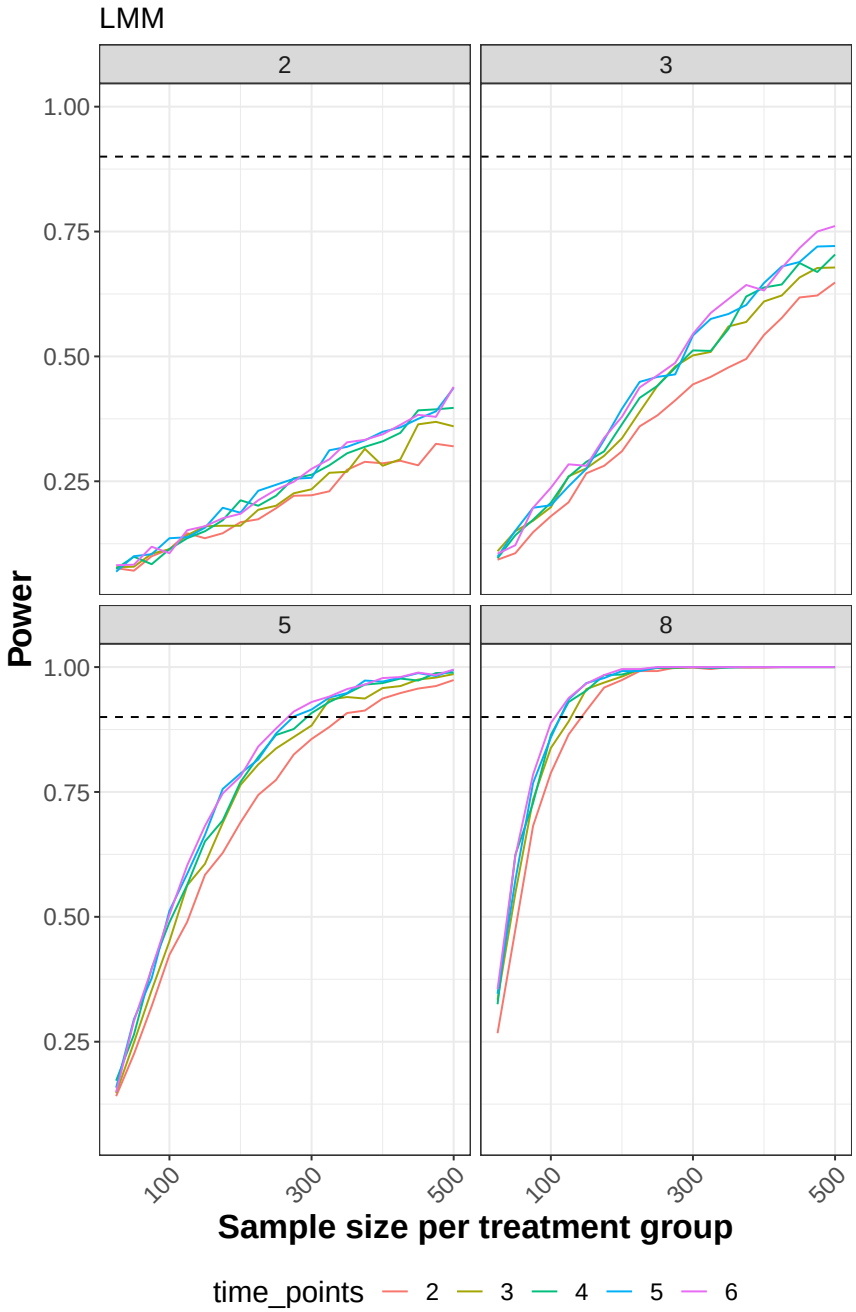

Figure 5.9: Power analysis based on simulations using the LMM and differing numbers of sampled time points with the CRD values of 2, 3, 5 and 8 QoL units (each CRD represented by an individual plot). The ICC was set at 0.5. The dashed line represents a power level of 0.9.

According to equation 3.17 if correlation (ρ) between within-patient measurements is less than one, increasing the number of sampled time points per patient will

decrease the required sample size, given other parameters are fixed. Put another way, increasing the number of sampled time points per patient will result in greater power for a fixed sample size. This is observed in figure 5.9. For CRDs of 3,5 and 8 QoL units, there appears a relatively greater difference in power between simulations utilising 2 and 3 sampled time points, as compared to simulations utilising 3 and 4, 4 and 5 and 5 and 6 time points. Power curves which crossed the 0.9 power threshold were observed using CRD of 5 and 8 QoL units. As the variation of these two sets of power curves appeared largest using CRD of 5 QoL units, this value of CRD was utilised to compare different models.

Simulations using the AR, exchangeable, independent GEE models and the LMM are displayed in figure 5.10. Simulations for 2 time points were technically not possible to perform for the unstructured GEE model. The simulations for the exchangeable, independent and LMM models appear similar. Differences between the 2 and 3 time points power curves for the AR model appear slightly smaller compared to the other models, although this may be due to sampling simulation error.

From examination of figure 5.11 the increase in power between 2 and 3 sampled time points and between 3 and 4 sampled time points for all sample sizes displayed is similar for all models. The increase in power generated from increasing sampled time points from 4 to 5 and from 5 to 6 appears to be minimal.

Examination of figure 5.10 demonstrates an apparently relatively large difference in power between 2 and 3 sampled time points at sample sizes between 200 and 250. The power values at these levels were between approximately 0.65 and 0.85. Table 5.2 depicts simulated power values using models with a sample size of 250 and a power around 0.8, and using models with a sample size of 325 and a power around 0.9. At the sample size of 250, all simulations using all models demonstrated a power increase of between 0.037 (exchangeable) and 0.063 by increasing sampled time points from 2 to 3. by using these three models, increasing sampled time points from 3 to 4, 4 to 5 or 5 to 6 resulted in a power change of less than 0.031. Utilising a sample size of 325, and by increasing sampled time points from 2 to 3, the increase in power using the LMM and exchangeable model was 0.055 and 0.058 respectively. Increasing the sampled time points 3 to 4, 4 to 5 and 5 to 6 produced a maximum increase in power of 0.01. Using the AR model, power changes for all single incremental increases in sample sizes were between 0.01 and 0.02. Little change in power occurred between time points using the independent model, except an increase of 0.04 between sampled time points 3 to 4. This is likely due to simulation sampling error. From table 5.2 there is evidence that increasing sampled time points from 2 to 3 does increase the power of the model around nominal power values of 0.8 and 0.9. At this level the effect of increasing the number of sampled time points above 3 appears limited.

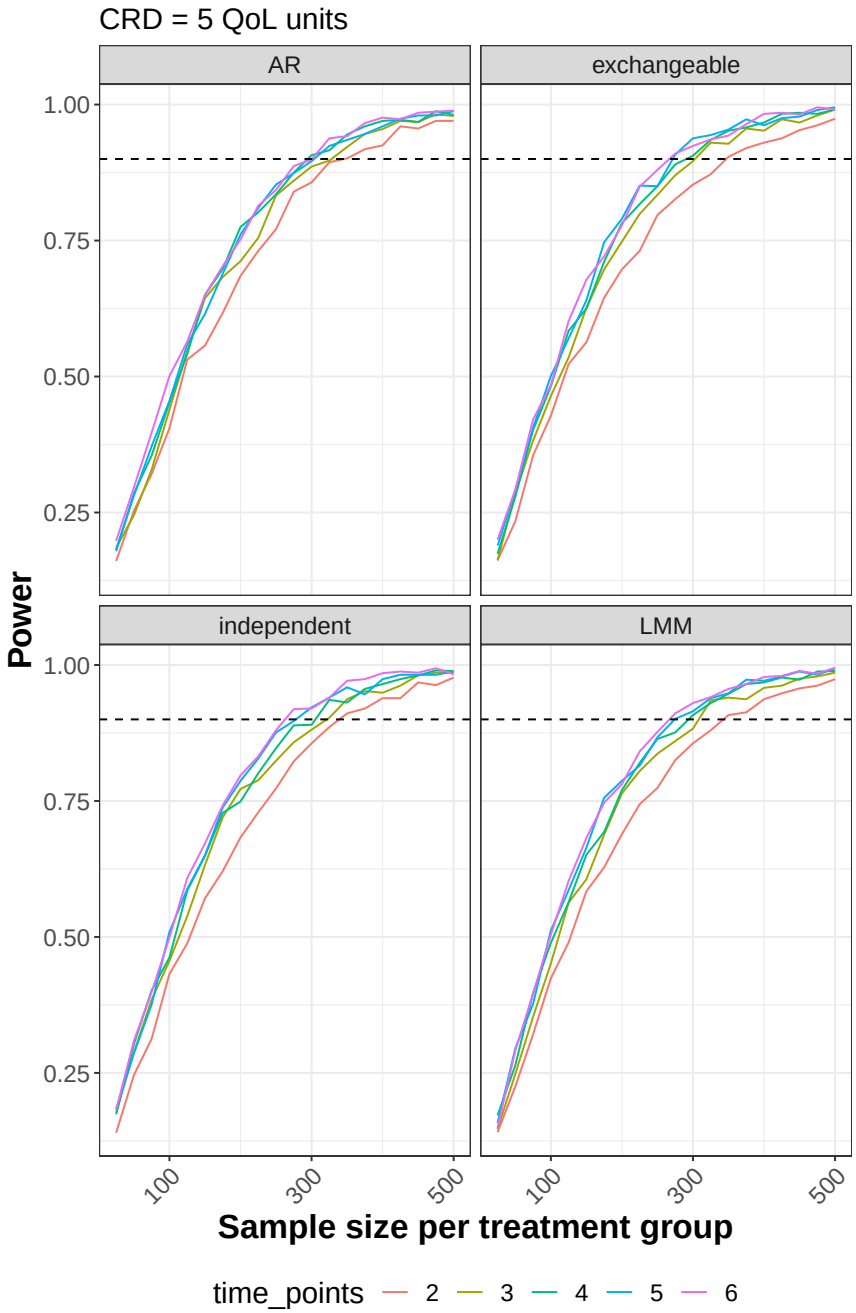

Figure 5.10: Power analysis based on simulations using LMM and the GEE AR, exchangeable and independent models with a CRD of 5 QoL units. Between 2 to 6 sampled time points were utilised in the simulations. The ICC was set at 0.5. The dashed line represents a power level of 0.9.

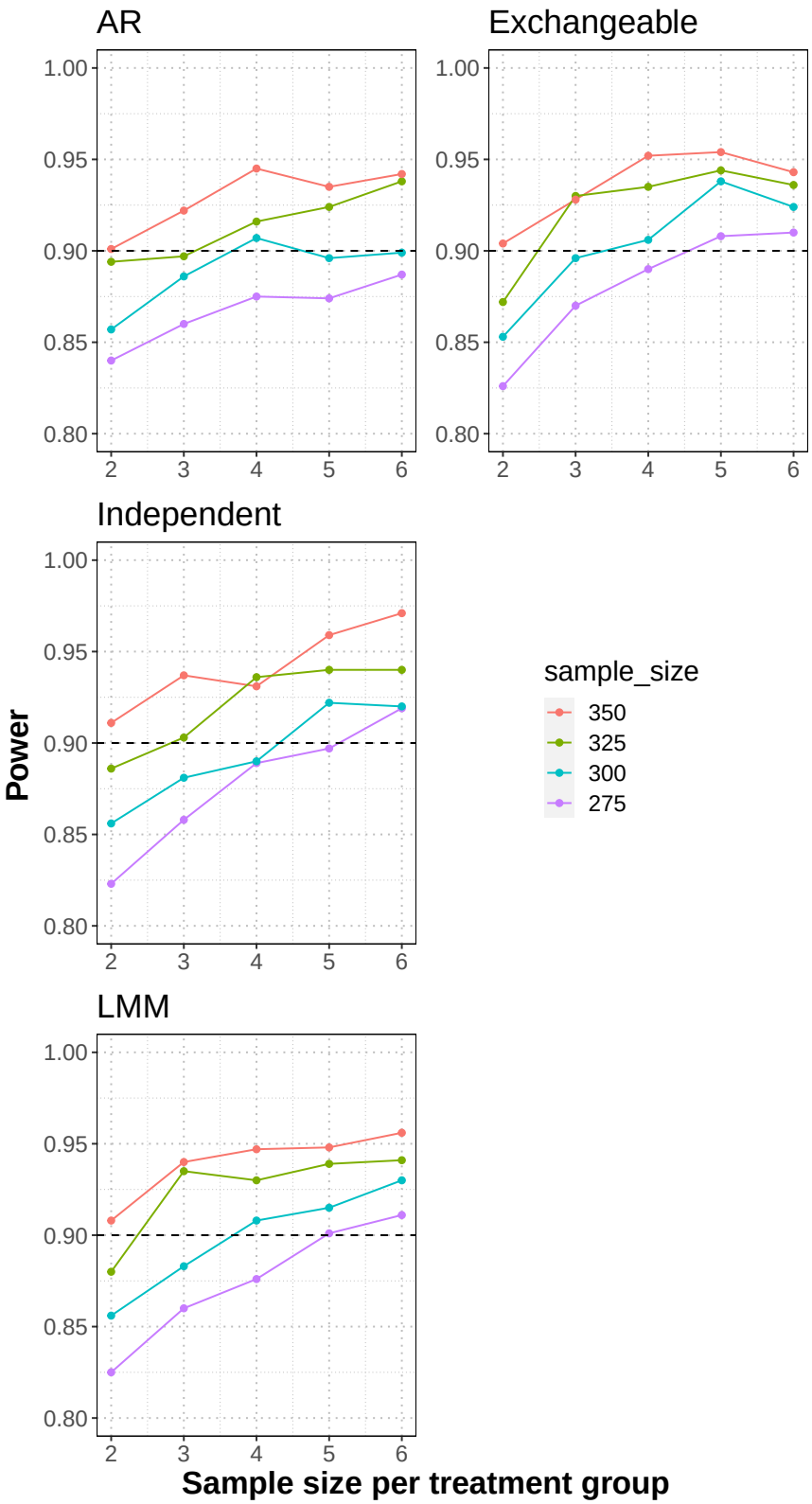

Figure 5.11: Rate of change of power using models containing 2, 3, 4, 5 and 6 numbers of sampled time points at 350, 325, 300 and 275 sample sizes per treatment group. The CRD value was set at 5 QoL units. The dashed line represents the 0.9 nominal power.

Table 5.2: Power analysis performed via simulation utilising sample sizes that approximate the nominal 0.8 (m=250) and 0.9 (m=325) levels and a range of sampled time points. (AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured)

		Models			
sample_size	time_points	AR	EX	IND	LMM
250	2	0.771	0.797	0.773	0.774
	3	0.833	0.834	0.824	0.837
	4	0.835	0.850	0.847	0.864
	5	0.853	0.850	0.876	0.867
	6	0.843	0.880	0.880	0.877
325	2	0.894	0.872	0.886	0.880
	3	0.897	0.930	0.903	0.935
	4	0.916	0.935	0.936	0.930
	5	0.924	0.944	0.940	0.939
	6	0.938	0.936	0.940	0.941

The simulated data generated for this report was based on an LMM with a random intercept and residual variances. The covariance matrix of the exchangeable model is the marginalised equivalent of the random intercept LMM variance. Zhang and Ahn (2011) calculated the reduction in sample size required for the exchangeable model with $\rho = 0.5$. Increasing observed measurements from 2 to 3 resulted in a 8.33% reduction and increasing observed measurements per patient from 3 to 4 resulted in a 4.17% sample size reduction. Unfortunately simulations in this report did not permit the reduction in sample sizes between the amount of sampled time points to be evaluated. However figures 5.9, 5.10 and 5.11 indicated a reduction in power benefit with increasing number of observations per patient.

The question of increasing the number of time points to reduce the required sample size is important. Increasing the number of repeated measures per patient may be cost-effective as compared to recruiting more patients. Yet a study design should be guided by both substantive as well as statistical concerns (Petras 2016). From a substantive point of view, the study must be designed with enough time points to capture time-specific dynamics. Results of simulations in this report suggest that a greater increase in power occurs when increasing sampled time points from 2 to 3 compared to higher combinations of sampled sizes. This finding is in keeping with the results published by Zhang and Ahn (2011).

5.8 Missing data

A common method of sample size estimation in RCTs with repeated measurements to accommodate missing data is by the formula

$$m = \frac{m_0}{1 - q} \tag{5.1}$$

where m_0 is the sample size estimate assuming no missing measurements and q is the dropout rate. Using such an approach may lead to overestimation of the required sample size (Zhang and Ahn 2012).

The missing data pattern utilised in this report (section 4.5.1.2) consisted of a 10% dropout rate at the 2nd post-baseline visit and 20% dropout rate at the third post-baseline visit. The relationship between simulations using complete data sets and missing value data sets for the LMM are displayed in figure 5.12. A comparison of complete and missing data for the GEE models resulted in similar patterns to the LMM (not shown). From figure 5.12 there appears evidence that the presence of the missing data does reduced the power of the simulation, albeit by a minimal amount. For low CRD of 2 QoL units the difference in power between the complete cases and missing data was sometimes negligible, such that greater power was observed from simulations incorporating missing data rather than simulations based on complete cases. At a CRD of 8 QoL units, the difference in power resulting from simulations incorporating missing data and complete cases was difficult to discern. At a CRD of 5 QoL units this difference was discernible .

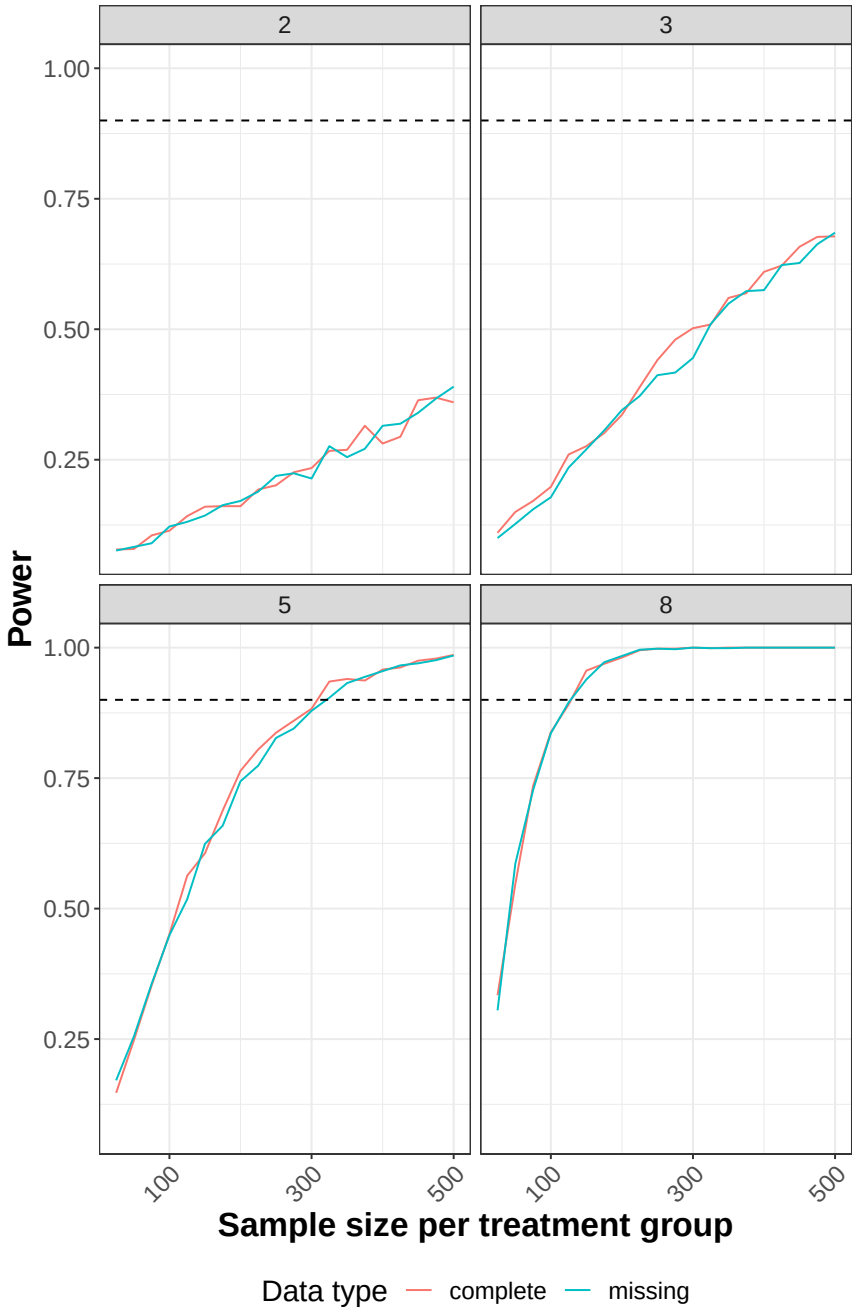

Figure 5.12: Power analysis of complete cases and missing data with LMM. Each plot represents power analysis using a CRD of 2,3,5 or 8 QoL units. ICC was set at 0.5. At 12 months, 10 percent of the sample size for each simulation was randomly assigned a missing value. At 24 months, this value was increased to 20 percent. Each column represents simulations performed on the CRD. The dashed line represents a power level of 0.9.

The power resulting from simulations using all models at a CRD of 5 QoL units is displayed in table 5.3. For a sample size of 250 per treatment arm, the power

Table 5.3: Power analysis performed via simulation using complete cases and missing data, an ICC=0.5 and a CRD of 5 QoL units. (AR = autoregressive, EX = exchangeable, IND = independent, LMM = liner mixed model, UN = unstructured)

		Models				
sample_size	data_type	AR	EX	IND	LMM	UN
225	complete	0.755	0.799	0.788	0.805	0.773
	missing	0.756	0.773	0.762	0.774	0.797
250	complete	0.833	0.834	0.824	0.837	0.849
	missing	0.807	0.841	0.811	0.827	0.813
275	complete	0.860	0.870	0.858	0.860	0.868
	missing	0.842	0.837	0.859	0.845	0.850
325	complete	0.897	0.930	0.903	0.935	0.914
	missing	0.874	0.901	0.904	0.904	0.912
350	complete	0.922	0.928	0.937	0.940	0.936
	missing	0.907	0.904	0.922	0.932	0.922

generated from simulations using all the 5 models for complete cases approximated 0.835. Using the exchangeable model, an increase in power incorporating missing data was recorded relative to complete cases. This is an indication of the minor effect on power of missingness in the simulations. For the AR, independent, LMM and unstructured models reductions in power due to missingness were recorded as 0.026, 0.013, 0.010 and 0.036 respectively. Using a sample size of 325 per treatment arm (approximating a power of 0.9) there was a negligible difference in power between complete and missing cases using the independent and unstructured model. For the AR, LMM and exchangeable models there was a reduction in power for missing data of 0.023, 0.031 and 0.029, respectively. The fact that power is generally biased downwards but not upwards for the missing data pattern indicates that missing data does reduce the power of simulations. However, the reduction in power, which varied from 0 to 0.03, indicates the missing data pattern in this report had only a minor effect on overall power.

In order to estimate the magnitude of sample size increase due to the missing data pattern investigated in this report, plots of the power analysis using all models at the 90% power level were created 5.13. Sample sizes estimated from the plots for complete cases and missing data respectively were 328 and 346 for the AR model, 302 and 320 for the exchangeable model, 322 for both cases for the independent model, 306 and 320 for the LMM and 306 and 318 for the unstructured model. Using equation 5.1, the maximum calculated dropout rate based on plots within figure 5.13 was 0.06 for the exchangeable model. This value is less than the 0.1 dropout rate at the 2nd post-baseline visit and considerable less than 0.2 rate at the 3rd post-baseline visit. According to simulations in this report, utilising the common method (equation 5.1) to account for missing data would overestimate the sample size required. This is in agreement with

the findings of Zhang and Ahn (2012). However it must be emphasised that missing data in this report was generated using the MCAR mechanism. The assumption of MCAR may not be valid in many clinical trials. Therefore, as part of a simulation study it is important to account for the missing data mechanism as well as the missing data pattern when generating simulated data sets.

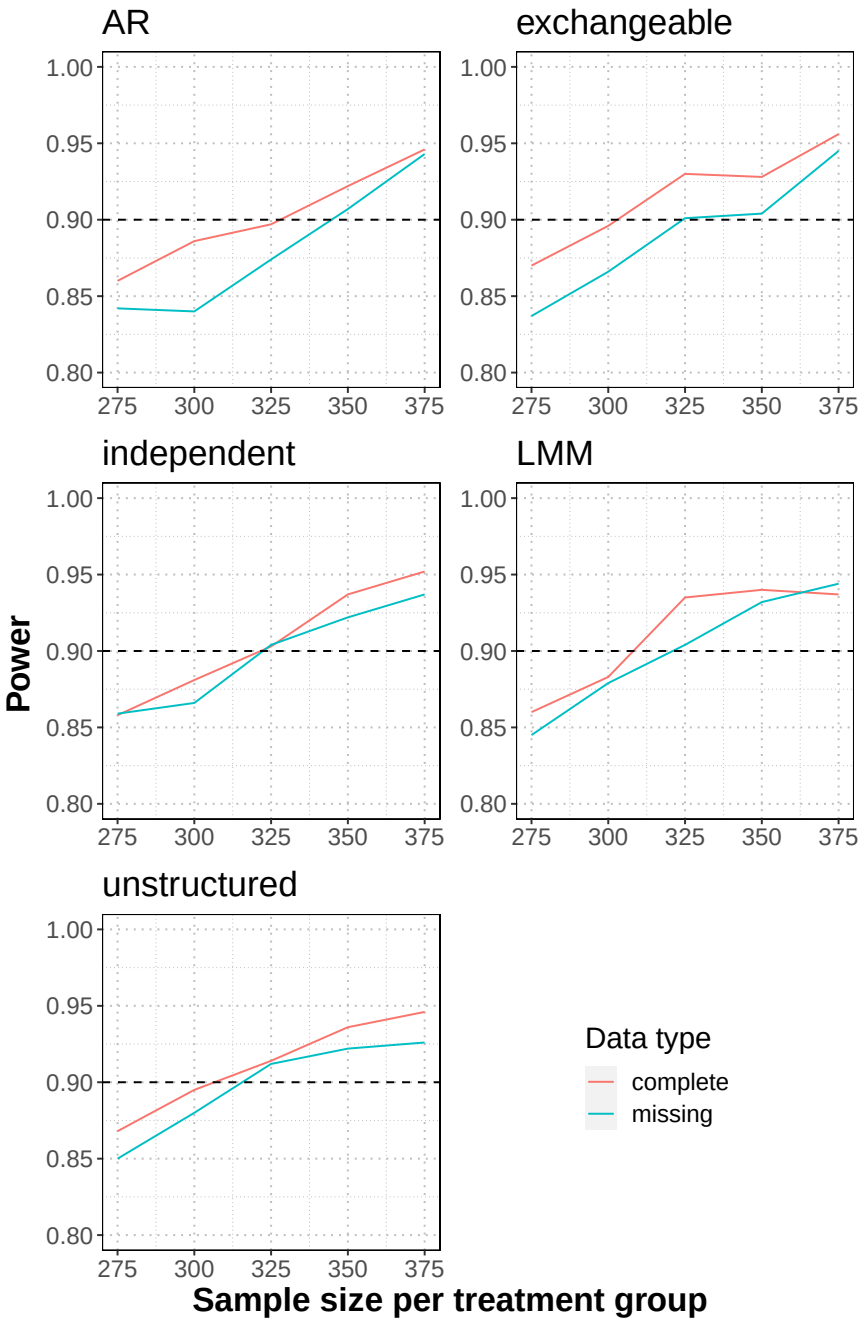

Figure 5.13: Power analysis of complete cases and missing data using all models, an ICC=0.5 and a CRD of 5 QoL units. The dashed horizontal line represents the 0.9 power level.

5.9 Covariance matrix misspecification

For sufficiently large sample sizes GEE produce consistent estimators for the marginal model equations. A very appealing property of GEE models is that for sufficiently large sample sizes, specification of working correlation matrix is not necessary. By utilising the *sandwich estimator*, even when the working correlation matrix has been misspecified, in many cases valid standard errors can be obtained using GEE (Fitzmaurice, Laird, and Ware 20012, p358). Throughout this report, the specification of the correlation matrix had a minimal observable effect on the power simulation. This result is in keeping with the view that GEE accommodate well misspecifications of the correlation matrix.

5.10 Limitations and recommendations

Effects of the different parameters on power investigated in this study were based on the dataset of Thomas et al. (2006). No interaction between time and treatment group were observed in the data, thereby restricting analysis to TAD between the acupuncture and usual care treatment groups. Analysis of a dataset using LMM that contained an interaction effect between time and group effect would be useful, in order to determine parameters for simulation studies. Extending this concept further, it would be worthwhile obtaining a dataset from which an LMM model could be developed that contained the aforementioned interaction, as well as a random slope effect. Simulation studies could then be performed to measure the effect of altering factors discussed in section 5.1 on power using LMM and GEE models.

The effect of unbalanced sample sizes as well as a a change in variance $Var(y_{ij})$ at time point $j = 1, ..., n_i$ on power could be explored.

Apart from the time and group covariate the only other significant covariate in the LMM model of the dataset supplied by Thomas et al. (2006) was a baseline pre-randomised measurement. It would be worthwhile obtaining a dataset that contains several significant covariates and examine their effect on power using LMM and GEE models.

The dataset supplied by Thomas et al. (2006) contained only a continuous outcome response. The effect of non-continuous outcome responses on power estimation has not be explored theoretically nor practically in this report. For non-continuous outcome responses, the population-averaged coefficients of GEE models are usually smaller than the equivalent fixed-effects coefficients of the LMM (Agresti 2007, p301). It is therefore recommended that datasets containing binary and count outcomes be examined, followed by relevant power analysis studies, in order to determine the outcome response effect on power using LMM and GEE models.

Chapter 6

Conclusions

A review of the literature was performed on the topic of sample size estimation in RCTs using repeated measures designs. Modeling methods were found to be popular in analysing longitudinal data. Two commonly utilised methods described in the literature and utilised in this study were linear mixed modeling (LMM) and marginal models analysed by generalised estimating equations (GEE). LMMs are designed to model patient-specific effects whereas GEEs are designed to measure the population-averaged effects. Approximation sample size formulas for LMM models were found in the literature.

The dataset from Thomas et al. (2006) was modeled using a LMM. As no evidence of a difference in the change of regression line slopes between the two treatment groups was found, the appropriate modeling problem became the estimation of time-average differences (TAD) between the two groups. The acupuncture interventional group was found to have an average value of 6.1 QoL units greater than the usual-care control group.

Power curves were generated from hypothesis testing of nested models with and without the group variable using 1000 simulations. Utilising the parameters of the LMM of the Thomas et al. (2006) dataset , inflated type I errors were observed for all models at very low sample sizes. Unsurprisingly these errors were greater for hypothesis testing using GEE models than for the LMM. The appearances of power curves generated from the hypothesis testing of LMM and GEE models at an $ICC = 0.5$ using simulated data were very similar to that generated from the TAD sample size formula 3.17. This suggests that the approximate sample size formula is precise for the dataset supplied by Thomas et al. (2006), and that little difference in power estimation exists when utilising hypothesis testing for any of the 5 models.

All results derived from simulations within this study were expected based on the current body of literature. The effect of increasing correlation for assessment of TAD between two treatment groups reduced power for all models and counteracted the effect of increasing CRD. Increasing the number of repeated

measures within the simulation resulted in a reduced increase in power as the number of measures was increased. The commonly utilised formula to account for missing data (equation 5.1) was found to overestimate the sample size required for repeated measures designs. For all hypothesis testing no obvious differences in the generated power curves were observed between any of the 5 models.

Appendix A

R Code for Model Specification

```
#function to detect columns containing time dependent variables ----
select_cols <- function(search_string){
          df %>%
          colnames() %>%
          str_detect(search_string) %>%
          which()
}

#load stata data into R dataframe
path = file.path("AcupunctureRCT.sav")
df <- read_sav(path)

#create tibble with relevant fields
#rename treatment field to group, and change coding to words
df <- df %>%
  select("studyid", "randomisation", "kbaspain", "k3pain",
         "k12pain", "k24pain") %>%
  mutate(group = (if_else(randomisation ==1, "acupuncture", "usual_care"))) %>%
  select(-c(randomisation)) %>%
  select(group, everything())

#remove all records that contain NA values for the fields k3/12/24pain
df <- df %>%
  filter(!is.na(k3pain)| !is.na(k12pain)| !is.na(k24pain))

#convert wide format (3,12,24 months) into long post-baseline format

    #select (3,12,24 months) columns
    col_nums <- select_cols(paste0("^k[1-9]*pain$"))
```

```
#convert into long post-baseline format
    pain_df <- df %>%
        gather(col_nums, key="time", value = "value") %>%
        select("studyid", "group", "kbaspain", "time", "value") %>%
        rename("response" := value) %>%
         mutate(time =
                case_when(
                                  time == "k3pain" ~ 3,
                                  time == "k12pain" ~ 12,
                                  time == "k24pain" ~ 24
                                ))


#check NA entries for each field

pain_df %>%
  group_by(group) %>%
  tally()

pain_df %>%
  group_by(group) %>%
  filter(!is.na(response)) %>%
  summarise(mean(response))

#convert group and studyid as factors and have baseline group as "usual_care"
pain_df <- pain_df %>%
  mutate( studyid=factor(studyid), group = factor(group)) %>%
  mutate(group = relevel(group, "usual_care"))

#remove NA cells in pain_df
pain_df <- pain_df %>%
  filter(!is.na(response))
```

Scaling the data

```
#Offset time data so that (3,12, 24 months) becomes (0,9,21 months)
#relative to 3 month date.
#Scale data so that response scale (0-1) is similar to
#the study time scale (0-1.75).

 scaled_df <- pain_df %>%
    mutate( response = response/100, kbaspain = kbaspain/100,
            time = (time-3)/12, group, studyid)
```

Which random effects can be dropped from the model?

```
#Can we drop slope and covariances random effects?
#slope, covariance and intercept RE
model_1fullslope <- lmer(response ~ kbaspain + time * group +
                           (time|studyid), data =scaled_df, REML = TRUE)

model_1int <- lmer(response ~ kbaspain + time * group +
                     (1|studyid), data =scaled_df, REML = TRUE)

ll_slope_cor <- logLik(model_1fullslope) #loglikelihood slope-intercept model
ll_int <- logLik(model_1int)          #loglikelihood intercept model

LLslopeint <- as.numeric(-2*(ll_int-ll_slope_cor))   #GLRT test statistic

1-pchisq(LLslopeint,2) #p=0.37  drop slope and covariances.
                               #keep intercept RE only
```

Which fixed effects can be dropped from the model?

```
  #Fixed effects
#do we need  interaction term group*time?
model_2 <- lmer(response ~ kbaspain + time * group   + (1|studyid),
                data =scaled_df, REML = FALSE)

model_3 <- lmer(response ~ kbaspain + time + group   + (1|studyid),
                data =scaled_df, REML=FALSE)

#parametric bootstrapping
N<-200
boot.test.stats<-rep(0,N)
set.seed(123)
for(i in 1:N){
  new.y<-unlist(simulate(model_3))
  fm.reduced.new<-lmer(new.y~kbaspain + time+group  + (1|studyid),
                       REML=F, data=scaled_df)
  fm.full.new<-lmer(new.y~kbaspain + time*group + (1|studyid),
                    REML=F, data=scaled_df)
  boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
}

mean(boot.test.stats> -2*(logLik(model_3) - logLik(model_2)))
#pvalue = 0.285 - no interaction term required

#Approximate chi-squared distribution
anova(model_2, model_3) #p-value = 0.32
```

```
#Is the time term necessary?
model_3a <- lmer(response ~ kbaspain + group + time   + (1|studyid),
                 data =scaled_df, REML=FALSE)

model_4 <- lmer(response ~  kbaspain + group + (1|studyid),
                 data =scaled_df, REML = FALSE)

N<-200
boot.test.stats<-rep(0,N)
set.seed(123)
for(i in 1:N){
new.y<-unlist(simulate(model_4))
fm.reduced.new<-lmer(new.y~ kbaspain + group + (1|studyid),
                     REML=F, data=scaled_df)
fm.full.new<-lmer(new.y~kbaspain + group+time + (1|studyid),
                  REML=F, data=scaled_df)
boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
}

mean(boot.test.stats> -2*(logLik(model_4) - logLik(model_3a))) #pvalue = 0.015
#keep time term

anova(model_3a, model_4) # pvalue=0.003


#do we need group term?

model_3b <- lmer(response ~  kbaspain + time + group + (1|studyid),
                 data =scaled_df, REML = FALSE)
model_5 <- lmer(response ~ kbaspain + time + (1|studyid),
                 data =scaled_df, REML = FALSE)

N<-200
boot.test.stats<-rep(0,N)
set.seed(123)
for(i in 1:N){
  new.y<-unlist(simulate(model_5))
  fm.reduced.new<-lmer(new.y~ kbaspain + time   + (1|studyid),
                       REML=F, data=scaled_df)
  fm.full.new<-lmer(new.y~kbaspain + time + group + (1|studyid),
                    REML=F, data=scaled_df)
  boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
}
```

```
mean(boot.test.stats> -2*(logLik(model_5) - logLik(model_3b))) #pvalue = 0.035
#keep group term

anova(model_3b, model_5) #pvalue = 0.026

#do we need baseline term?

model_3c <- lmer(response ~   time + group + kbaspain +(1|studyid),
                 data =scaled_df, REML = FALSE)
model_6 <- lmer(response ~ time + group + (1|studyid), scaled_df)

boot.test.stats<-rep(0,N)
set.seed(123)
for(i in 1:N){
  new.y<-unlist(simulate(model_6))
  fm.reduced.new<-lmer(new.y~group +time  + (1|studyid),
                       REML=F, data=scaled_df)
  fm.full.new<-lmer(new.y~group+time   + kbaspain + (1|studyid),
                    REML=F, data=scaled_df)
  boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
}

mean(boot.test.stats> -2*(logLik(model_6) - logLik(model_3c))) #pvalue = 0.00
# keep baseline term

anova(model_3c, model_6) # p=0.00000

#model_3 is chosen for extraction of parameters for simulations
summary(model_3)
```

Appendix B

Model checking

The random effects model expressed in equation 4.1 follows the formula

$$y_{ij} = (\beta_0 + b_{0i}) + \beta_{baselinse} + \beta_{time} \times time_{ij} + \beta_{treatment} \times treatment + \epsilon_{ij},$$
$$i = 1, .., M, \quad j = 1, ..., n_i$$
$$b_{0i} \sim N(0, \sigma_a^2), \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

where β_0, $\beta_{baselinse}$, β_{time} and $\beta_{treatment}$ represent the fixed effects coefficients for the intercept, baseline, time and treatment variables respectively. The explanatory variable *treatment* was given a value 1 for patients in the acupuncture group and 0 for patients in the usual care group.

The model is based on the following assumptions:

- Assumption 1: the random effects b_i are i.i.d $N(0, \sigma_b^2)$.
- Assumption 2: the within-subject (residual) errors ϵ_{ij} are i.i.d $N(0, \sigma^2)$ and independent of the random effects b_i.
- Assumption 3: the fixed effects β_0, $\beta_{baselinse}$, β_{time} and$\beta_{treatment}$ adequately represent the mean response $E(y_{ij})$.

Assessment of assumption 1 was performed by examining figure B.1

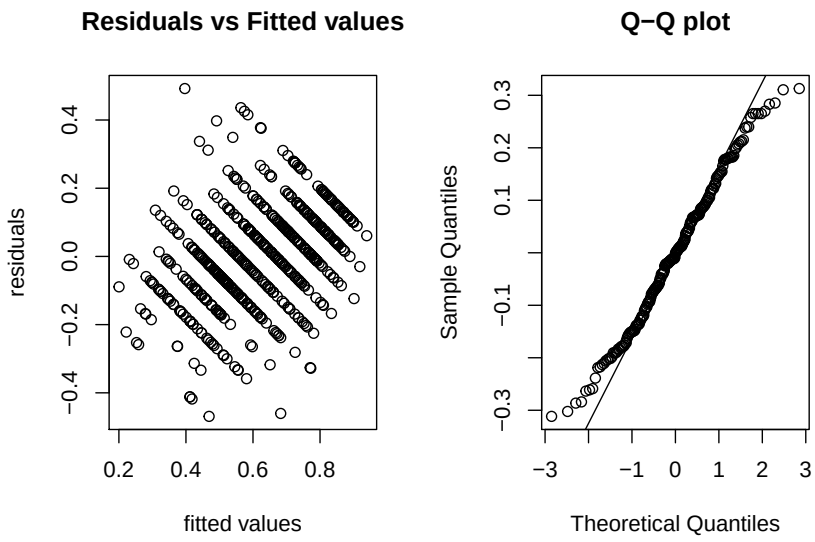

Figure B.1: Diagnostic plots for between-patient random effects

The plot of fitted vs residual values for between-subject errors should demonstrate random scatter. Because the response variable had only 10 possible values it is difficult to interpret the scatter pattern. There appears insufficient evidence to claim that random scatter has not occurred. The normal plot demonstrates a good degree of normality, with only slight deviation at the tails. Therefore it is plausible to accept assumption 1.

Assumption 2 is assessed by investigating the within-subject residual errors. The diagnostic plots are shown in figure B.2

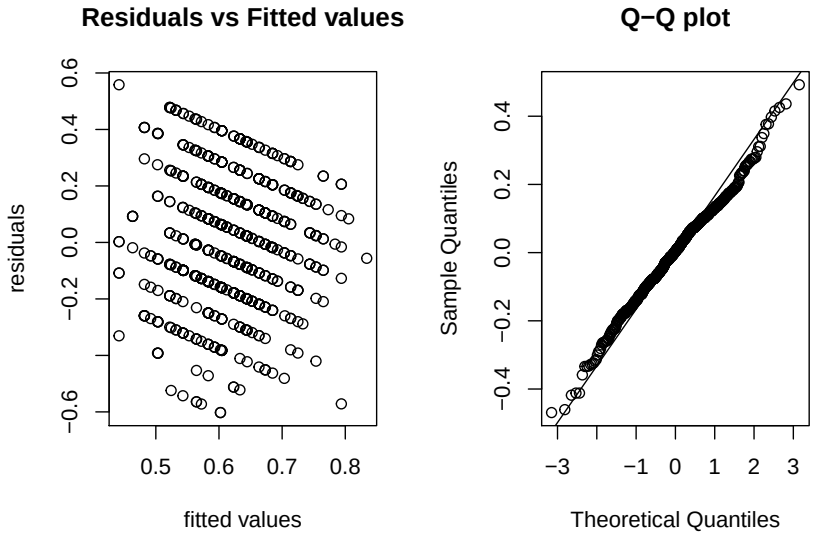

Figure B.2: Diagnostic plots for within-subject errors

In the Q-Q plot of figure B.2 all the within-subject residual errors lie close to the straight line, indicating that the data are normally distributed. As for the within-subject variation, the residual vs fitted values plot is difficult to interpret due to the limited range of values for the response variable. However, there doesn't appear enough evidence to suggest that the residuals are not randomly distributed. For this reason,
it is plausible to accept assumption 2.

Finally, assumption 3 can be assessed by looking at the relationship between the actual vs fitted response outcomes, seen in figure B.3.

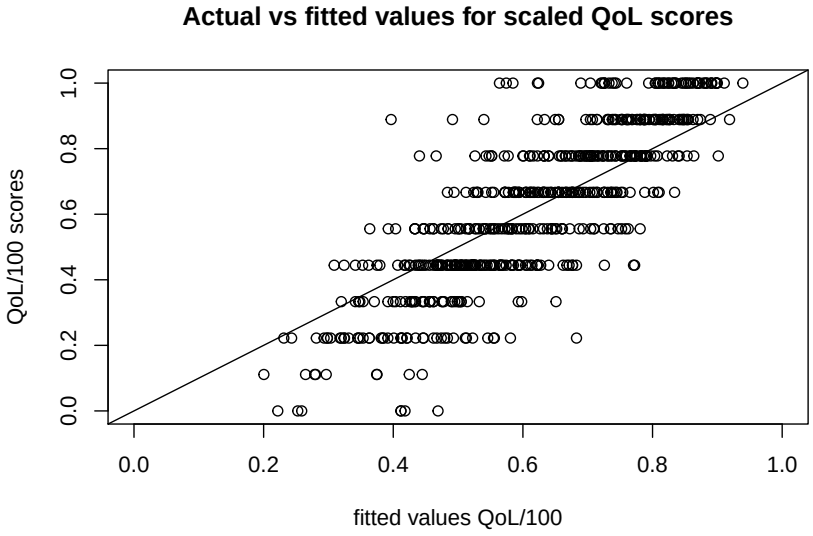

Figure B.3: Relationship between actual and fitted values for the scaled QoL score

Ideally, the data should vary around the straight line with unit slope (as seen in figure B.3). It is evident that the data vary around a straight line but not around the unit slope. This discrepancy may well be due to the discrete nature of the response variable. Therefore it is credible to accept assumption 3.

As all the above assumptions have been accepted, the random effects model proposed in chapter 5 for the data of Thomas et al. (2006) is considered valid.

Appendix C

Simulation R code

C.1 Complete cases data simulations

C.1.1 Power function

This function creates a set number of simulated datasets. Each dataset is modeled with and without the treatment variable, and a GLRT applied using a χ^2_1 distribution to determine if the difference between models is significant at $p = 0.05$ level. The number of significant tests is totaled and expressed as a ratio of total simulated datasets. The ratio is the power of the test.

The following parameters were defined a-priori:

- G (number of treatment groups)
- beta_sigma (the between-subject variance)
- sigma (the within-subject variance)

Baseline values for simulations were selected randomly with replacement from the relevant treatment group in the dataset supplied.

```
library(haven)
library(dplyr)
library(geepack)
library(lme4)

#mixed model function for power determination

power.f <- function(N, group_diff, n, t_points){
  #N : amount of simulations
  #group_diff : the difference in QoL between the acupuncture
  #n : sample size of each group
  #t_points : the time points for post-baseline repeated measures

  p.stat<-rep(0,N)  #vector that stores
```

```
for(i in 1:N){

  #creates group, subject, kbaspain(basline) and  beta values in df
test.df <- data.frame(group = factor(rep(1:G, each=n),
                                     labels = c("usual_care", "acupuncture") ),
                      studyid = as.factor(c(1:(n*G))),
                      beta_i =  rnorm(n*G, 0, beta_sigma))

  #sample kbaspain for each group from original dataset with replacement
  kb_usual_care <- df %>%
    filter(group == "usual_care") %>%
    select(kbaspain) %>%
    pull() %>%
    sample(n, replace=T)

  kb_acu <- df %>%
    filter(group == "acupuncture") %>%
    select(kbaspain) %>%
    pull() %>%
    sample(n, replace=T)

  test.df <- test.df %>%
    mutate(kbaspain = c(kb_usual_care, kb_acu))

  #replicate each above entry by the number of time points
  #utilised in the simulation
  test.df <- replicate(length(t_points), test.df, simplify = F) %>%
    bind_rows() %>%
    arrange(studyid)

  #add field with time point values eg 0,0.75 and 1,75
  #add field with within-patient noise
  test.df <- test.df %>%
    mutate(time = rep(t_points, 2*n)) %>%
    mutate(within_var = rep(rnorm(n*G*length(t_points),0,sigma)))

  #calculation of response value
  # adding group_diff value to response of acupuncture group
  test.df <- test.df %>%
    mutate(response = intercept + beta_i + time*time_slope + within_var) %>%
    mutate(response = if_else(group =="acupuncture",
                              response+group_diff, response))

  #create full model with group var and create reduced model
```

```
    #without group var using simulated data
    full.model <- lmer(response~  kbaspain +  time + group +
                         (1|studyid), data=test.df, REML = F )
    reduced.model <- lmer(response~  kbaspain + time +
                            (1|studyid), data=test.df, REML = F )

    #calculate pvalue from GLRT
    p.stat[i] <- anova(full.model, reduced.model)[[2,8]]

  }

  #calculate proportion of simulations with pvalues below 0.05 to
  #determine power of test
  length(p.stat[p.stat<0.05])/length(p.stat)
}
```

In order to examine the effect of GEE models (eg for exchangeable covariance pattern), the following code is substituted for the full and reduced models

```
#create full model with group var and create reduced model
    #without group var using simulated data
    #GEE model with exchangeable covariance pattern
full.model <- geeglm(response~  kbaspain +  time + group,
                     id=studyid, data=test.df,
                     family="gaussian", corstr="ex")
reduced.model <- geeglm(response~  kbaspain +  time,
                     id=studyid, data=test.df,
                     family="gaussian", corstr="ex")
```

C.1.1.1 Parametric bootstrapping

The p-value for the GLRT according to parametric bootstrapping was determined by substituting the code

```
 #calculate pvalue from LRT
    N_boot<-200
    boot.test.stats<-rep(0,N_boot)
    for(j in 1:N_boot){
      new.y<-unlist(simulate(reduced.model))
      fm.reduced.new<-lmer(new.y~time + kbaspain + (1|studyid),
                           data=test.df, REML = F )
      fm.full.new<-lmer(new.y~group + time + kbaspain + (1|studyid),
                        data=test.df, REML = F )
      boot.test.stats[j]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))

    }
```

```
    p.stat[i] <- mean(boot.test.stats> -2*(logLik(reduced.model) -
                                          logLik(full.model)))
```

for that found in the previous section

```
    #calculate pvalue from GLRT
    p.stat[i] <- anova(full.model, reduced.model)[[2,8]]
```

C.1.2 Calculating power curve

The power function above is incorporated into another function that calculates the power for a range of sample sizes and treatment differences.

```
library(haven)
library(dplyr)
library(geepack)
library(lme4)

#--------variables for simulation------------
time_points <- c(0, 9/12, 21/12) # years from 3 months post-baseline
treat_diff <- c(0,0.02, 0.03, 0.05, 0.08) #QoL/100
sample_size <- c(seq(25, 500, 25))


#------------1st stage in selecting kbaspain values from original data-----------
path = file.path("AcupunctureRCT.sav")
df <- read_sav(path)

df <- df %>%
  select("randomisation", "kbaspain") %>%
  rename(group = "randomisation") %>%
  mutate(group = if_else(group ==1, "acupuncture", "usual_care"))

#------ parameters obtained from original data
#data obtained from regression of original dataset
intercept <- 4.417e-01  #intercept fixed effect obtained
time_slope <- 0.0287    #slope fixed effect
sigma <- 0.1661         #within subject variance
var_sigma <- sigma^2
var_beta_sigma <-  var_sigma # random-intercept variance with
                             #interclass correlation (ICC) = 0.5
beta_sigma <- sqrt(var_beta_sigma)  #between-subject sd
G <- 2                  #number of groups

options(contrasts =c("contr.treatment", "contr. poly") )
```

```
#-------simulation---------------

#testing for estimates of power
treat_diff <- c(0,0.02, 0.03, 0.05, 0.08)
sample_size <- c(seq(25, 500, 25))

sim.results <- function(sim.size, pow.func, file.name){
  obj.name <- matrix(rep(NA, length(sample_size)*length(treat_diff)),
                     nrow=length(sample_size) , ncol=length(treat_diff))

  colnames(obj.name) <- c(paste0(treat_diff*100))
  rownames(obj.name) <- c(paste0(sample_size))

  #simulation loop
  for (i in 1:length(sample_size)) {
    for(j in 1:length(treat_diff)){
      obj.name[i,j] <- pow.func(sim.size, treat_diff[j],
                                sample_size[i], time_points)
    }
  }
#saves data into external R datafile
  save(obj.name, file = paste0(file.name, ".RData"))
}

sim.results(1000, power.f, "output_file")
```

C.2 Missing data simulations

C.2.1 Power function

Missing data was simulated according to the MCAR pattern. For a 3 time point post-baseline simulation (3,12,24 months), the dataframe was split into three separate lists. Missing data points were randomly sampled for each time point and given the value *NA*.

```
#mixed model function for power determination
power.f.dropoff <- function(N, group_diff, n, t_points, dropoff){
  p.stat<-rep(0,N)  #vector that stores

  for(i in 1:N){

    #creates group, subject, kbaspain and  alpha values in df
    test.df <- data.frame(group = factor(rep(1:G, each=n),
                                  labels = c("usual_care", "acupuncture") ),
```

```
                        studyid = as.factor(c(1:(n*G))),
                        beta_i =  rnorm(n*G, 0, beta_sigma))

#sample kbaspain for each group from original dataset with replacement
kb_usual_care <- df %>%
  filter(group == "usual_care") %>%
  select(kbaspain) %>%
  pull() %>%
  sample(n, replace=T)

kb_acu <- df %>%
  filter(group == "acupuncture") %>%
  select(kbaspain) %>%
  pull() %>%
  sample(n, replace=T)

test.df <- test.df %>%
  mutate(kbaspain = c(kb_usual_care, kb_acu))

# in this case three sub-dataframes are created (length(time_points)=3)
temp.df <- replicate(length(time_points), test.df, simplify = F)

#for each sub-dataframe eg  test.df[[2]],
#the number of missing data entries eg.dropoff[1]
# are randomly sampled and replaced with NAs.
tmp.tp2 <- as.numeric(sample(test.df[[2]],
                             round(n*2*dropoff[1]), replace=F) )
temp.df[[2]][tmp.tp2,3] <- NA

tmp.tp3 <- as.numeric(sample(test.df[[2]],
                             round(n*2*dropoff[2]), replace=F) )
temp.df[[3]][tmp.tp3,3] <- NA

test.df <- temp.df %>%
  bind_rows() %>%
  arrange(studyid)

#add field with time values 0 and 1.75
#add field with within noise
test.df <- test.df %>%
  mutate(time = rep(time_points, 2*n)) %>%
  mutate(within_var = rep(rnorm(n*G*length(time_points),0,sigma)))

#calculation of response value
# adding group_diff value to response of acupuncture group
```

```
    test.df <- test.df %>%
      mutate(response = intercept + beta_i + time*time_slope + within_var) %>%
      mutate(response = if_else(group =="acupuncture",
                                response+group_diff, response)) %>%
      filter(!is.na(beta_i) )

   #create full model with group var and create reduced model
    #without group var using simulated data
    full.model <- lmer(response~  kbaspain +  time + group +
                         (1|studyid), data=test.df, REML = F )
    reduced.model <- lmer(response~  kbaspain + time +
                            (1|studyid), data=test.df, REML = F )

    #calculate pvalue from LRT
    p.stat[i] <- anova(full.model, reduced.model)[[2,8]]

  }

  #calculate proportion of simulations with pvalues below 0.05 to determine po
  length(p.stat[p.stat<0.05])/length(p.stat)

}
```

C.2.2 Calculating the power curve for missing data

This script is similar to that stated previously, with the exception that the percentage of MCAR pattern is denoted for the 2nd and 3rd time points ie. dropoff <- c(0.1,0.2).

```
library(haven)
library(tidyverse)
library(geepack)
library(lme4)

#--------variables for simulation------------
time_points <- c(0, 9/12, 21/12) # years from 3 months post-baseline
treat_diff <- c(0,0.02, 0.03, 0.05, 0.08) #QoL/100
sample_size <- c(seq(25, 500, 25))

#dropoff ratio for 2nd and 3rd time point respectively eg 1 = all NA values,
#0.5 = half of total acupuncture and usual_care responses = NA
dropoff <- c(0.1,0.2)

#------------1st stage in selecting baseline values from original data--------
path = file.path("AcupunctureRCT.sav")
```

```r
df <- read_sav(path)

df <- df %>%
  select("randomisation", "kbaspain") %>%
  rename(group = "randomisation") %>%
  mutate(group = if_else(group ==1, "acupuncture", "usual_care"))

#------ parameters obtained from original data
#data obtained from regression of original dataset
intercept <- 4.417e-01  #intercept fixed effect obtained
time_slope <- 0.0287    #slope fixed effect
sigma <- 0.1661         #within subject variance
var_sigma <- sigma^2
var_beta_sigma <-  var_sigma # random-intercept variance with
                             #interclass correlation (ICC) = 0.5
beta_sigma <- sqrt(var_beta_sigma)  #between-subject sd
G <- 2                  #number of groups

options(contrasts =c("contr.treatment", "contr. poly") )

#-------simulation---------------

sim.results <- function(sim.size, pow.func, file.name){
  obj.name <- matrix(rep(NA, length(sample_size)*length(treat_diff)),
                     nrow=length(sample_size) , ncol=length(treat_diff))

  colnames(obj.name) <- c(paste0(treat_diff*100))
  rownames(obj.name) <- c(paste0(sample_size))

  #simulation loop
  for (i in 1:length(sample_size)) {
    for(j in 1:length(treat_diff)){
      obj.name[i,j] <- pow.func(sim.size, treat_diff[j], sample_size[i],
                                time_points, dropoff)
    }
  }

  save(obj.name, file = paste0(file.name, ".RData"))
}

sim.results(1000, power.f.dropoff, "output_file")
```

References

Agresti, Alan. 2007. *An introduction to categorical data analysis*. 2nd edn. New Jersey: John Wiley & Sons Ltd.

Ahn, Chul, and Sin Ho Jung. 2005. "Effect of dropouts on sample size estimates for test on trends across repeated measurements." *Journal of Biopharmaceutical Statistics* 15 (1): 33–41.

Bahçecitapar, M. K. 2018. "Some factors affecting statistical power of approximate tests in the linear mixed model for longitudinal data." *Communications in Statistics: Simulation and Computation* 47 (1): 294–314. https://doi.org/10.1080/03610918.2017.1283699.

Brandmaier, A. M., T. von Oertzen, P. Ghisletta, U. Lindenberger, and C. Hertzog. 2018. "Precision, reliability, and effect size of slope variance in latent growth curve models: Implications for statistical power analysis." *Frontiers in Psychology* 9 (APR). https://doi.org/10.3389/fpsyg.2018.00294.

Diggle, Peter J., Patrick J. Heagerty, Kung-Yee Liang, and Scott L. Zeger. 2002. *Analysis of Longitudinal Data*. 2nd edn. Oxford: Oxford University Press.

Fitzmaurice, Garrett M, Nan M Laird, and James H Ware. 20012. *Applied Longitudinal Analysis*. New Jersey: John Wiley & Sons Ltd.

Frison, L., and S. J. Pocock. 1992. "Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design." *Statistics in Medicine* 11 (13): 1685–1704. https://doi.org/10.1002/sim.4780111304.

Guo, Yi, Henrietta L Logan, Deborah H Glueck, and Keith E Muller. 2013. "Selecting a sample size for studies with repeated measures." *BMC Medical Research Methodology* 13 (1): 100. https://doi.org/10.1186/1471-2288-13-100.

Guthrie, Bruce, Shaun Treweek, Dennis Petrie, Karen Barnett, Lewis D Ritchie, Chris Robertson, and Marion Bennie. 2012. "Protocol for the Effective Feedback to Improve Primary Care Prescribing Safety (EFIPPS) study : a cluster randomised controlled trial using ePrescribing data," 1–9. https://doi.org/10.1136/bmjopen-2012-002359.

Hayes, Richard J., and Lawrence H. Moulton. 2016. *Cluster Randomised Statistical Analysis of Shapes Astrostatistics Bayesian Disease Mapping : Epidemiology*

Statistics in Clinical Pharmacology Clinical Trials in Oncology Second Edition Cluster Randomised Trials Correspondence Analysis Design and Analysis. Boca Raton: CRC Press.

Hedeker, Donald, and Robert D. Gibbons. 2006. *Longitudinal Data Analysis*. New Jersey: Wiley-Interscience.

Johnson, Paul C. D., Sarah J. E. Barry, Heather M. Ferguson, and Pie Müller. 2015. "Power analysis for generalized linear mixed models in ecology and evolution." *Methods in Ecology and Evolution* 6 (2): 133–42.

Julious, Steve, and N A. 2010. *Sample sizes for clinical trials*. Chapman & Hall/CRC.

Kain, Morgan P., Ben M. Bolker, and Michael W. McCoy. 2015. "A practical guide and power analysis for GLMMs: Detecting among treatment variation in randomeffects." *PeerJ* 2015 (9). https://doi.org/10.7717/peerj.1226.

Lane, S. P., and E. P. Hennes. 2019. "Conducting sensitivity analyses to identify and buffer power vulnerabilities in studies examining substance use over time." *Addictive Behaviors* 94: 117–23. https://doi.org/10.1016/j.addbeh.2018.09.017.

Liang, Kung-Yee, and Scott L Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

Liu, Guanghan, and Kung-Yee Liang. 1997. "Sample Size Calculations for Studies with Correlated Observations." *Biometrics* 53 (3): 937–47.

Lou, Ying, Jing Cao, and Chul Ahn. 2017. "Sample size estimation for comparing rates of change in K-group repeated count outcomes." *Communications in Statistics - Theory and Methods* 46 (22): 11204–13. https://doi.org/10.1080/03610926.2016.1260744.

Lou, Ying, Jing Cao, Song Zhang, and Chul Ahn. 2017a. "Sample size calculations for time-averaged difference of longitudinal binary outcomes." *Communications in Statistics—Theory and Methods* 46 (1): 344–53.

———. 2017b. "Sample size estimation for a two-group comparison of repeated count outcomes using GEE." *Communications in Statistics - Theory and Methods* 46 (14): 6743–53. https://doi.org/10.1080/03610926.2015.1134572.

Magnusson, K, G Andersson, and P Carlbring. 2018. "The consequences of ignoring therapist effects in trials with longitudinal data: A simulation study." *Journal of Consulting and Clinical Psychology* 86 (9): 711–25. https://doi.org/10.1037/ccp0000333.

Matthews, J N S, Douglas Altman, M J Campbell, and P Royston. 1990. "Analysis of serial measurements in medical research." *British Medical Journal* 300: 230–35.

McNeish, Daniel M., and Jeffery R. Harring. 2017. "Clustered data with small sample sizes: Comparing the performance of model-based and design-based

approaches." *Communications in Statistics: Simulation and Computation* 46 (2): 855–69.

Moerbeek, Mirjam, and Steven Teerenstra. 2015. *Power Analysis of Trials with Multilevel Data*. https://doi.org/10.1201/b18676.

Morel, J. G., M. C. Bokossa, and N. K. Neerchal. 2003. "Small sample correction for the variance of GEE estimators." *Biometrical Journal* 45 (4): 395–409.

Morrel, C Jane, Pauline Slade, Rachel Warner, Graham Paley, Simon Dixon, Stephen J Walters, Traolach Brugha, Michael Barkham, Gareth J Parr, and Jon Nicholl. 2009. "Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care." *British Medical Journal* 338: 1–12.

Overall, John E., and Suzanne R. Doyle. 1994. "Estimating sample sizes for repeated measurement designs." *Controlled Clinical Trials* 15 (2): 100–123.

Overall, John E., and Robert R. Starbuck. 1979. "Sample size estimation for randomized pre-post designs." *Journal of Psychiatric Research* 15 (1): 51–55.

Petras, H. 2016. "Longitudinal Assessment Design and Statistical Power for Detecting an Intervention Impact." *Prevention Science* 17 (7): 819–29.

Piaggio, Gilda, Diana R Elbourne, Stuart J Pocock, Stephen J W Evans, and Douglas G Altman. 2012. "Reporting of noninferiority and equivalence randomized trials: Extension of the CONSORT 2010 statement." *JAMA - Journal of the American Medical Association* 308 (24): 2594–2604.

Pinheiro, Jose C, and Douglas M Bates. 2000. *Mixed Effects Models in S and S-Plus*. New York: Springer-Verlag.

Skene, S S, and M G Kenward. 2010. "The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator." *Statistics in Medicine* 29 (27): 2825–37. https://doi.org/10.1002/sim.4073.

Tango, T. 2017. *Repeated measures design with generalized linear mixed models for randomized controlled trials*. Boca Raton: CRC Press. https://doi.org/10.1201/9781315152097.

Thomas, K. J., H. MacPherson, L. Thorpe, J. Brazier, M. Fitter, M. J. Campbell, M. Roman, S. J. Walters, and J. Nicholl. 2006. "Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain." *British Medical Journal* 333 (7569): 623–26. https://doi.org/10.1136/bmj.38878.907361.7C.

Tu, Xin M., J. Kowalski, P. Crits-Christoph, and R. Gallop. 2006. "Power analyses for correlations from clustered study designs." *Statistics in Medicine* 25 (15): 2587–2606. https://doi.org/10.1002/sim.2273.

Walters, Stephen J. 2009. *Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation*. https://doi.org/10.1002/9780470840481.

Wang, Jijia, Song Zhang, and Chul Ahn. 2020. "Sample size estimation for comparing rates of change in K-group repeated binary measurements studies." *Communications in Statistics - Theory and Methods* 0 (0): 1–10. https://doi.org/10.1080/03610926.2020.1736302.

Zhang, S, and C Ahn. 2012. "Sample Size Calculation for Time-Averaged Differences in the Presence of Missing Data." *Contemp Clin Trials* 33 (3): 550–56.

Zhang, Song, and Chul Ahn. 2011. "How many measurements for time-averaged differences in repeated measurement studies ?" *Contemporary Clinical Trials* 32 (3): 412–17. https://doi.org/10.1016/j.cct.2011.01.002.