The analysis of multiple correlated outcome measures in randomised controlled trials

Victoria Vickerstaff

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of the

University College London (UCL).

Declaration

I, Victoria Vickerstaff, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

……..

Victoria Vickerstaff

Abstract

Multiple primary outcomes are sometimes collected and analysed in randomised controlled trials (RCTs), and are used in favour of a single outcome. By collecting multiple primary outcomes, it is possible to fully evaluate the effect that an intervention has for a given disease process. A simple approach to analysing multiple outcomes is to consider each outcome separately, however, this approach does not account for any pairwise correlations between the outcomes. Any cases with missing values must be ignored, unless an additional imputation step is performed. Alternatively, multivariate methods that explicitly model the pairwise correlations between the outcomes may be more efficient when some of the outcomes have missing values.

When analysing multiple outcomes in a trial, it is important to control the family wise error rate (FWER), which is the probability of finding at least one false positive result. A common approach is to adjust the p-values for each statistical test. It is also important to consider the power to detect the true effects of the intervention.

In this thesis, I present an overview of the relevant methods that could be used to analyse multiple outcomes in RCTs, including methods based on multivariate multilevel models. I perform simulation studies to provide guidance on which methods should be used to adjust for multiple comparisons in the sample size calculation, and which methods should be used for the analysis when the multiple primary outcomes are correlated. Additionally, I use simulation studies to investigate the differences in the power obtained when using multivariate models compared to analysing the outcomes separately using univariate models. Different simulation scenarios were constructed by varying the number of outcomes, the type of outcomes, the degree of correlations between the outcomes and the proportions and mechanisms of missing data.

Impact Statement

Randomised controlled trials (RCTs) are the most rigorous way to investigate the effectiveness of a new intervention. The conclusions drawn from RCTs can provide evidence that can be used to decide whether an intervention should become routinely available for patients.

Statistical methods are necessary for the design of trials and analysis of trial data. The use of appropriate statistical tests is essential to prevent errors and biases, and therefore the reporting of erroneous conclusions in medical research. Even though appropriate statistical methods are needed to ensure only effective interventions become available, often inappropriate statistical tests are used and hence potentially unreliable results are reported. The aim of the work presented in this thesis is to provide guidance and recommendations for the design and analysis of RCTs that use multiple outcomes.

When designing an RCT and analysing trial data, it is necessary to consider the number of outcomes involved. I reviewed trials published in high impact clinical journals and have shown that methods accounting for multiple outcomes are not regularly used when calculating the sample size or when analysing trial data. My statistical investigations have shown that different conclusions may have been drawn in certain published trials if the correct steps had been taken to account for multiple comparisons during the analysis of trial data. The number of incorrect analyses observed in recently published RCTs demonstrates that there is a current need for clear guidance for the design and analysis of RCTs that use multiple outcomes.

One of the practical recommendations I have provided is how to determine a sufficient sample size. The sample size is an important consideration as the number of trial participants is restricted by economic, ethical and practical considerations. On the other hand, if the sample size is too small it may not be possible to correctly determine whether an intervention works. I have described several approaches that can be used to determine the required sample size for trials involving multiple outcomes. I have discussed that the chosen approach would depend on the clinical objective of the trial. For example, the objective might be to ascertain whether an intervention is effective on at least one of the outcomes. Alternatively, it could be ascertain whether an intervention is effective on all of the outcomes. I have shown that the sample size required varies depending on the chosen clinical objective.

The research performed in this thesis has the potential to have a wide impact as it is directly relevant to researchers who work on RCTs. Some of the work in this thesis has already been published and cited (Vickerstaff et al., 2015, Vickerstaff et al., 2019). Should researchers follow the guidance provided, it is expected that their analyses could potentially be more robust in that there is a higher chance that only successful interventions are identified as such; and more efficient in that fewer patients are recruited for RCTs with multiple outcomes.

Acknowledgements

I am grateful to my supervisors Prof. Rumana Z. Omar and Dr. Gareth Ambler for their help, guidance, patience and support. I have benefited immensely from their expertise and encouragement. This PhD would not have been possible without their encouragement to begin the journey of a part-time PhD. Throughout the PhD, they have provided so much support which has kept me motivated and allowed me to develop as a researcher. I am very fortunate to have been encouraged to pursue not only a PhD, but also a career, in an area that interests me greatly.

I have also been very fortunate to work with Prof. Michael King and Prof. Irwin Nazareth who have been generous with their advice and support. I would also like to thank the teams within Marie Curie Palliative Care Research Department (MCPCRD) and The Research Department of Primary Care and Population Health (PCPH). In particular, I would like to thank all my colleagues within the MCPCRD. They have provided so much support both academically and emotionally, without them, I would have been lost.

I would like to express my gratitude to the Department of Statistical Science, MCPCRD and PCPH for funding my PhD research. Their support has allowed me to complete this PhD part-time alongside my academic position. Additionally, I would like to thank Dr Rebecca Beeken, Dr. Marta Buszewicz, and Prof. Sonia Johnson for providing me with data that I have used as motivating examples throughout my PhD.

Finally, I would like to thank my family. My Mum has set an amazing example and showed me I could do anything, be anything. Her support and continual encouragement throughout life have made me who I am today. To my husband, Louis, thank you for joining me on all my adventures and supporting me throughout this process.

Table of Contents

List of Tables

List of Figures

Abbreviations

BDI-II	Beck Depression Inventory
BMC	BioMed Central
CI	Confidence interval
CONSORT	Consolidated Standards of Reporting Trials
COPD	Chronic Obstructive Pulmonary Disease
CPMP	Committee for Proprietary Medicinal Products
D/AP	Dubey/Armitage-Parmar
EmpSE	Empirical Standard Error
EQ-VAS	EuroQol-visual analogue scales
ES	Effect size
FCS	Fully conditional specification
FWER	Familywise error rate
GP	General Practice
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
IQR	Interquartile range
JAMA	Journal of the American Medical Association
LV	Latent variable
MANOVA	Multivariate analysis of variance
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MCSE	Monte Carlo standard error
MI	Multiple imputation
MICE	Multiple imputation by chained equations
MM	Multivariate multilevel model
MNAR	Missing not at random
MSE	Mean square error
N	Number
NA	Not applicable
NEJM	New England Journal of Medicine
PAAS	Prospective alpha allocation scheme

PANSS	Positive and Negative Syndrome Scale
RCT	Randomised controlled trial
RSA	R2-adjustment methods
SD	Standard deviation
SE	Standard error
TCH	Tukey-Ciminera-Heyse method
UK	United Kingdom
UV	Univariate model
WHO	World Health Organization
WSAS	Work and social activities scale

Chapter 1

Introduction

Randomised controlled trials (RCTs) can be used to investigate the effectiveness of a new intervention. An RCT is a study in which people are randomly assigned to two (or more) groups to test the effect of a specific intervention on a health outcome of interest. In a standard two-arm trial, one group receives the intervention being tested and the other group receives a dummy intervention (placebo) or the usual care. The trial participants are followed over time and their outcome data are collected to assess the effect of the new intervention. In most clinical trials a single primary outcome is specified to investigate the effect of a health intervention and this is often sufficient to determine whether the intervention is effective. However, for many diseases and disorders, a patient's health status cannot be adequately quantified using a single primary outcome. Examples include mental health disorders, stroke (Mayo and Scott, 2011) and chronic obstructive pulmonary disease (COPD) (Agusti and Vestbo, 2011, Teixeira-Pinto et al., 2009, De Los Reyes et al., 2011). Therefore, in these disease areas, multiple primary outcomes may be required to provide a comprehensive understanding of the effects of an intervention.

In trials, multiple statistical tests may be performed to investigate the effect of the intervention when analysing multiple outcome measures. Each time a statistical test is performed, there is a chance that a statistically significant effect will be observed due to chance when no effect is present. This is known as a '*type I error'*. As the number of statistical tests performed on the same dataset increases, the probability of a type I error increases. The issue associated with the increased chance of a type I error is referred to as '*multiplicity'*. It is essential that multiplicity is accounted for when designing and analysing RCTs. Another problem when analysing outcome data in RCTs is the failure to detect a true effect of the intervention. This is known as a `*type II e*rror'. The power of a study is defined as one minus the probability of a type II error. The desired power of the study is usually specified a priori. A sample size calculation is typically performed in a trial to ensure that sufficient number of participants are recruited to achieve the desired power. The sample size calculation is an important part of designing a clinical trial to ensure that the trial is efficient, ethical and cost effective (Röhrig et al., 2010).

Several methods have been proposed in the literature to address the issue of multiplicity, however, many of these methods are not used in practice. In fact, many trials fail to account

for multiplicity in the design and analysis stages. For the practitioner, it is often unclear which (if any) of the proposed methods should be used to account for multiplicity whilst ensuring that the analysis remains efficient. It is also important that the statistical issues are communicated well to the clinicians to enable them to appropriately interpret the results.

To this end, the focus of this thesis is to evaluate the existing statistical techniques available for the design of RCTs and the analysis of trial data with multiple outcomes. In the remainder of this chapter I provide a brief introduction about how multiple outcomes may be analysed, followed by the aims, scope and structure of the remainder of the thesis.

1.1 Overview

Before a trial commences, the primary outcome measure(s) ('outcome(s)') need to be specified (WHO, 2012). The primary outcome can be defined as the outcome that is most capable of providing clinically relevant and convincing evidence that is directly related to the primary objective of the study (ICH E9 Expert Working Group, 1999). In this thesis, I focus on the statistical issues related to the use of multiple primary outcomes, as a single outcome is often not sufficient on its own to capture the range of clinically relevant intervention benefits for a particular health condition (European Medical Agency, 2017).

The multiple outcomes may have the same data type, for example, several continuous outcomes may be measured to quantify cognitive and behavioural components in order to evaluate the effect of cognitive behavioural therapy on patients with a depressive disorder. Alternatively, the outcomes may be of different data types. For example, researchers might measure a continuous quality of life outcome and a binary outcome to indicate symptom relapse when evaluating the effect of an antipsychotic drug on people with schizophrenia.

Missing outcome data is a common problem for RCTS since it is not always possible to measure all specified primary outcomes for all participants. In fact, a review of published trials showed that outcome data was missing in the majority of trials (Bell et al., 2014). Missing outcome data will generally results in a loss of power and may lead to biased estimated of the effect of the intervention., For example, patients in a smoking cessation trial may be more likely to drop out if they continue to smoke, and therefore the patients with observed outcome data may not be a representative samples.

Several approaches have been used to analyse trials with multiple outcomes in the presence of missing data. A commonly approach, which is appealing due to its simplicity, has been to

analyse the outcomes separately within a univariate framework (Pocock et al., 1987). Patients are typically omitted from any analysis for which they have missing outcome data. However, this approach does not account for the possible correlation between the outcomes and consequently the precision of the estimates and the power may be lower than that achieved by other approaches (Teixeira-Pinto et al., 2009). A variation on this approach is to use multiple imputation to impute missing outcome data prior to the univariate analysis of the outcomes (White et al., 2011). An advantage of this approach is that all outcomes may be included in the imputation model and hence the correlation between the outcomes may be accounted for (White et al., 2011).

More advanced methods include the use of multivariate methods such as the multivariate multilevel (MM) model and the multivariate regression. These multivariate methods have been used to analyse examination results in schools (Goldstein et al., 1993, Yang et al., 2002), crime trends (Mohan et al., 2011, Tseloni and Zarafonitou, 2008) and health-related behaviour (Maas et al., 2008). However, the use of these methods in trials has been limited despite their potential to increase power (Snijders and Bosker, 2012). For example, the MM has occasionally been used for an exploratory analysis in clinical trials (Hassiotis et al., 2009, King et al., 2002).

It is important to control for multiplicity in confirmatory trials, in which the goal of the trial is to confirm the effect of an intervention (Bender and Lange, 2001). It is crucial to ensure that correct inferences are made from these trials as they inform healthcare policy and medical practice.

It should be noted that the work in this thesis focuses on *multivariate* methods. The terms *multivariate* and *multivariable* are sometimes used interchangeably in the literature (Hidalgo and Goodman, 2013). However, these terms represent two types of analyses. A multivariable linear regression model is a model in which multiple covariates or 'independent' variables are used, for example, multiple covariates may be used to adjust the analyses for baseline factors or stratification factors. In contrast, a multivariate linear regression model has multiple outcomes or 'dependent' variables.

1.2 Aims and scope

The overall aim of the research is to address the need for an evaluation of methods to analyse multiple primary outcomes in clinical trials.

The aim is achieved by the following specific objectives:

- To review the literature of published RCTs to ascertain whether multiple primary outcomes are commonly used, and to identify the methods that are frequently used to account for multiplicity in the sample size calculation and analysis of RCTs.

- To evaluate the validity of existing methods that account for multiplicity arising from multiple primary outcomes. Multiplicity should be addressed both in the sample size calculation and statistical analysis. When using multiple primary outcomes, there is limited guidance as to which method(s) should be used to address multiplicity, especially when there are missing data in the primary outcomes. Using the results from several simulation studies, I will conclude on which methods may be used to account for multiplicity in the analysis of trials with multiple primary outcomes in the presence of missing data.

- To investigate the possibility of using multivariate models as opposed to univariate models for the analysis of RCTs with multiple outcomes, and to identify scenarios when multivariate methods may be advantageous. For clinical trials, it is important that the procedure has sufficient power to detect the effects of the intervention– when they are present – whilst controlling the type I error.

- To provide practical recommendations regarding the approaches to be used for the design and analysis of trials involving multiple outcomes.

1.3 Structure of the thesis

The remainder of this thesis is structured as follows. Chapter 2 is a review of statistical methods that are proposed in the literature to account for multiple outcomes. Chapter 3 is a review of RCTs recently published in high impact neurology and psychiatry journals to ascertain whether multiple outcomes are used in practice and how these outcomes are handled during the design and analysis of trials. Chapter 4 is a comparison of methods to

adjust p-values (or equivalently significance levels) to account for multiplicity in the sample size calculation and analysis of trials with multiple primary outcomes. Chapter 5 compares selected multivariate methods for the analysis of multiple outcomes in terms of type I error and power. Chapter 6 is a comparison of selected multivariate methods for the analysis of multiple outcomes in terms of bias associated with the estimated effects of the intervention. Chapter 7 investigates models that simultaneously analyse time-to-event and continuous outcomes. Chapter 8 provides discussion, guidance and conclusions.

Chapter 2

Background and key concepts

This chapter provides a concise summary of the background and key concepts that are required for the analysis of multiple outcomes. Definitions of the familywise error rate and power are provided in the context of multiple outcomes. The reasons for missing data can be classified as one of three 'missing data mechanisms'. Moreover, a number of methods that have been proposed to analyse multiple correlated outcomes in clinical trials are summarised. The topics covered provide a foundation for the subsequent chapters.

2.1 Notation

The following notation will be used throughout this thesis. The problem of multiplicity due to the assessment of multiple outcomes in RCTs is considered. This problem can be formulated in terms of m null hypotheses which are denoted by $H_{01}, \dots, H_{0m}$, respectively. Each null hypothesis corresponds to the assessment of a new intervention based on one of the m outcomes. The null hypothesis H_{0j} is defined in terms of a relevant intervention parameter θ_j, as given by

$$H_{0j}: \theta_j = \delta_j,$$

for $j = 1, \dots, m$. Here a value of θ_j greater than δ_j indicates a benficial effect and a value of θ_j less than δ_j indicates a harmful effect. For continuous outcomes it is usual to specify $\theta_j = \mu_{j1} - \mu_{j2}$ where μ_{j1} and μ_{j2} are the mean responses of the j^{th} outcome in the two intervention groups, respectively. In this setting, δ_j is usually 0. The null hypotheses are tested versus the alternative hypotheses, which are given by

$$H_{1j}: \theta_j \neq \delta_j,$$

for $j = 1, \dots, m$. $p_1, \dots, p_m$ denotes the marginal p-values for the appropriate statistical tests associated with $H_{01}, \dots H_{0m}$. Moreover, $p_{(1)}, \dots, p_{(m)}$ denotes the ordered p-values that correspond to the ordered null hypotheses that are denoted by $H_{0(1)}, \dots H_{0(m)}$, respectively.

2.2 Types of multiple primary outcomes

After specifying the primary outcome(s), investigators must identify a criterion to determine whether the intervention has demonstrated an effect. When performing an RCT with multiple primary outcomes, there are two main clinical decision rules that can be used to determine whether the intervention is a success: all primary outcomes need to be statistically significant; or, at least one of the primary outcomes needs to be statistically significant. The primary outcomes are known as co-primary outcomes and alternative outcomes, respectively.

2.2.1 Co-primary outcomes

The primary outcomes are known as co-primary outcomes when all outcomes must be statistically significant to show that the intervention is effective. In some instances, regulatory agencies have required that a statistically significant effect of the intervention is shown on several outcomes before deeming the intervention is effective (Offen et al., 2007). For example, for a regulatory agency to declare that a new migraine treatment is effective, the intervention needs to be shown to be effective on four outcomes: pain, nausea, photosensitivity and phonosensitivity.

For co-primary outcomes, it is recommended that all outcomes are tested at the same significance level, say 0.05 (Committee for Proprietary Medicinal Products, 2002). No adjustment for type I error is required to account for multiplicity when testing all co-primary outcomes at the same significance level, however, adjustments to the power to detect an intervention effect and sample size need to be considered. Depending on the correlations between the outcomes, there may be a large reduction in the power to detect an intervention effect (Offen et al., 2007). For example, for a trial with two independent co-primary outcomes, if the power to detect the desired effect is 80% for each outcome, then there would only be 64% (=80% x 80%) power to detect a true intervention effect on both outcomes. If the correlation between the two outcomes is 0.5, then the power to detect a true intervention effect is 69% (Offen et al., 2007).

2.2.2 Alternative outcomes

The primary outcomes are known as ‘alternative outcomes’ or ‘multiple primary outcomes’ when at least one outcome needs to be statistically significant in order to conclude that the intervention is effective. In this case, any of the pre-specified primary outcomes can, on their own, indicate a clinically meaningful benefit of the intervention. The benefits of the

intervention may be promoted differently depending on which outcome is shown to be statistically significant. When using alternative outcomes, adjustments are needed to control for inflated error rates (in particular, the familywise error rate defined below in Section 2.3). Many methods have been introduced to control the error rates when analysing multiple primary outcomes; these are described in Section 2.7. In this thesis, when discussing 'multiple outcomes', the focus is on the scenario of alternative outcomes.

2.3 Multiple comparisons theory

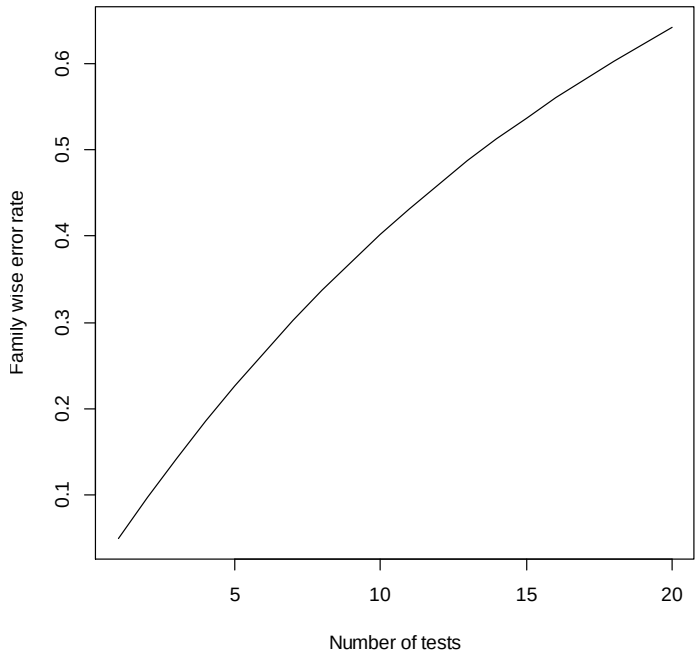
In this section, the definition of type I error and power is provided in the context of multiple outcomes.

2.3.1 Familywise Error Rate

Multiple comparisons must be performed when analysing multiple outcomes to investigate whether the intervention is effective for each outcome. If multiple comparisons are performed at the nominal significance level, then the overall probability of finding at least one false positive result can be unacceptably high. As a simple example, consider two outcomes that are analysed independently of each other and at the nominal significance level of 0.05. The probability of finding at least one false positive significant result is 0.098 ($= 1 - (1 - 0.05)^2$). This probability is greater than the nominal significance level and is known as the familywise error rate (FWER) (Alosh et al., 2014). The FWER obtained by analysing a varying number of outcomes independently is displayed in Figure 2.1 below. Due to the inflated FWER obtained when analysing multiple outcomes, it is important to account for the number of primary outcomes when performing the analyses for confirmatory RCTs.

When controlling the FWER, it is necessary to consider pairwise correlations between the outcomes (Phillips and Haudiquet, 2003). Selecting a method of analysis that ignores the correlations may lead to adjustments that are overly conservative. This could waste resources, as the required sample size is dependent on the method of primary analysis and will be larger than necessary. This would inflate the costs and duration of the study.

Figure 2.1 The familywise error rate obtained when analysing multiple outcomes without adjusting for multiplicity.



There are two approaches to controlling the FWER: strong control and weak control. The strong control is defined as the control of the probability of incorrectly rejecting any true hypothesis regardless of whether any of the other hypotheses are true (Dmitrienko et al., 2009). In other words, the strong control refers to the control of the type I error rate under any combination of true and false null hypotheses. It is necessary to have strong control of the FWER for the primary outcomes in all confirmatory clinical trials, as stated in the guidelines by regulators (Committee for Proprietary Medicinal Products, 2002). Weak control of the FWER is computed under the assumption that all of the hypotheses are true. Without any other safeguards, weak control of the FWER is unsatisfactory. Consequently, only methods that have strong control of the FWER are investigated.

There are other error rate definitions that apply when there are a large number of comparisons, for example, in genetic studies or microarray experiments, the false discovery rate has been used (Dmitrienko et al., 2009). These definitions are not commonly used in clinical trials and are therefore beyond the scope of this thesis.

2.3.2 Power

A key consideration for the design of clinical trials is the power of the trial to detect the effects of the intervention in question, when they are present. When there are multiple outcomes there are a number of ways to define the power. This is chosen depending on the

clinical objective of the study. Three definitions discussed here are: disjunctive power, conjunctive power and marginal power.

The disjunctive power (or minimal power), (Westfall et al., 2011), is the probability of finding at least one true intervention effect across all of the outcomes (Dmitrienko et al., 2009, Bretz et al., 2010). The conjunctive power (or maximal power) is the probability of finding a true intervention effect for all outcomes (Westfall et al., 2011). It may be noted that the disjunctive and conjunctive power have previous been referred to as 'multiple' and 'complete' power respectively (Westfall et al., 2011), however, this naming convention may lead to confusion since disjunctive power may be greater than the conjunctive power (Senn and Bretz, 2007). The marginal (or individual) power is the probability of finding a true intervention effect on a particular outcome. It is calculated separately for each outcome. When the clinical objective is to detect an intervention effect for at least one of the outcomes the disjunctive power and marginal power are recommended. The conjunctive power is recommended when the clinical objective is to detect an intervention effect on all the outcomes (Dmitrienko et al., 2009, Bretz et al., 2010).

2.4 Missing data theory

In this section, I introduce the possible 'missing data mechanisms' and discuss some methods that have been used in practice to handle missing data values in trials.

Missing data are observations that exist that could have been made but were not recorded, or were recorded but then lost. Almost all randomised trials have outcomes that have missing values as highlighted in a recent review of trials which found that 95% of trials reported some missing data (Bell et al., 2014). For clinical trials, missing data may result from the withdrawal of a participant or if a participant is lost to follow up. If the missing data are ignored or incorrectly handled then the conclusions drawn from the data could be incorrect (Carpenter and Kenward, 2007).

2.4.1 Missing data mechanisms

In this section, the missing data mechanisms are described. These mechanisms specify how the underlying value of the missing observation is associated with the reason for being missing. Rubin defines three missing data mechanisms: 'missing completely at random' (MCAR), 'missing at random' (MAR) and 'missing not at random' (MNAR) (Little and Rubin, 2014). These mechanisms are defined in turn below.

Missing completely at random

The outcome data are said to be MCAR if there are no systematic differences between the missing values and the observed values (Sterne et al., 2009). In this scenario, the missingness does not depend on the baseline covariates or the outcome. For example, for a weight loss trial if weight measurements are missing due to a malfunction of the weighing scales.

Missing at random

The outcome data are said to be MAR if the probability that the data are missing depends on the values of the observed data. However, conditional on the values of the observed data, the probability that the data are missing does not depend on the values of the missing data. In other words, any systematic difference between the missing values and the observed values can be explained by information in the observed data (Sterne et al., 2009). For example, continuing with the weight loss trial, assuming that the participants' age is observed and included in the trial analysis, older individuals are more likely to have their weight recorded by the GP but at any age, individuals with low and high weight are equally likely to have their weight recorded.

Missing not at random

The outcome variable is said to be MNAR if there are systematic differences between the missing values and the observed values, even after the information from the observed data is taken into account (Sterne et al., 2009). That is, the probability of a missing outcome depends on the unobserved outcomes as well as the observed data. Parameter estimation from the observed data alone is typically biased. The amount of bias depends on the extent of dropout and the strength of the relationship between the unobserved outcome and probability of dropout. For example, continuing the weight loss trial example, individuals who have gained weight may be more likely to miss appointments if they have not achieved their target weight loss.

There is no test that can identify whether the missing data mechanism is MCAR, MAR or MNAR. Consequently, when performing the analyses, it is necessary to make an assumption about the missing data. An incorrect assumption can lead to biased estimates, which can result in incorrect confidence intervals and, consequently incorrect p-values. Alternatively,

the analysis may be inefficient resulting in wider confidence intervals and larger p-values than necessary.

2.4.2 Methods of analysis with missing data

There are various methods that can handle missing data (Carpenter and Kenward, 2007). A few of the methods that have been used in published trials are described below (Wood et al., 2004).

Complete case analysis

A complete case analysis only considers the complete records. In other words, only the participants without missing values are included. When the missing data is MCAR this method results in unbiased estimates, however, the precision is reduced.

Multiple Imputation

Multiple imputation (MI) was first described by Rubin (Rubin, 1996). It follows from regression imputation (using the observed data to predict the missing values). The process is repeated numerous times to account for the uncertainty in the imputed values. The multiple imputation process can be broken down into three stages:

1) Imputation: numerous sets of plausible values are created to 'fill-in' the missing values to create 'complete' datasets.
2) Analysis: the desired analysis is performed on each of the complete data sets created in (1).
3) Pooling: the results from the repeated analyses are combined into a single result.

There are various methods available to perform the imputation step. A commonly used method is multiple imputation using chained equations (MICE). This is also known as Fully Conditional Specification (FCS) as each partially observed variable is imputed from its full conditional distribution given all other variables. MICE uses univariate models for each partially observed variable conditioned on all the other variables. If we had partially observed variables V_j for $j = 1, \ldots, m$, then the MICE method would create m univariate models. The algorithm is:

a) A simple imputation is performed, for example using the sample mean. All missing values are 'filled in'.

b) The imputations for V_1 are reset to missing.

c) V_1 is regressed against the other variables, and a regression equation is obtained.

d) The regression equation from stage c) is used to simulate the missing values in V_1.

e) Stages (b) to (d) are repeated for each $V_2, \ldots, V_m$ in turn. This is called a 'cycle'. At the end of each cycle all missing values in the dataset have been replaced with simulated values from regressions.

f) Stages (b) to (e) are then repeated so that a number of cycles are completed.

g) The final cycle in (f) provides a single imputed dataset. The process (a) to (f) is repeated a number of times to create multiple datasets which are then combined using Rubin's rules.

Alternatively, a Bayesian method can be used to perform the imputation step by sampling from the posterior distribution of the joint distribution for the missing data given the observed data. When the joint likelihood function is complex, and cannot be simulated from directly, Markov chain Monte Carlo (MCMC) may be used to obtain (approximate) simulated values.

2.5 Motivating examples

In this section, two clinical trials are described. These trials are revisited later in this chapter and subsequent chapters to illustrate the techniques. They are examples of real trials that analysed multiple correlated outcomes and motivate this work. The first trial collected data on three continuous outcomes. In contrast, the second trial collected a mixture of continuous and binary outcomes.

2.5.1 Pro-active Care and its Evaluation for Enduring Depression Trial, ProCEED

The ProCEED trial is a two arm, individually randomised controlled trial (Buszewicz et al., 2010, Buszewicz et al., 2016). The trial aims to establish whether structured and pro-active care of patients with chronic depression, in primary care, leads to a cost-effective improvement in medical and social outcomes when compared with the usual GP care over 24 months. The dataset includes a sample of 558 participants with chronic depression taken from 42 primary care practices across the United Kingdom.

The ProCEED trial used the Beck Depression Inventory (BDI-II), which is a measurement of severity of depression, as its primary outcome. The main results indicated that the practice

nurse-led proactive care was beneficial for some participants. However, the result for the primary outcome was not statistically significant at the 5% level (effect on BDI-II: 1.2 95% confidence interval (-0.3, 2.7) p=0.125) (Buszewicz et al. 2016). The trial team were also interested in the work and social activities scale (WSAS) and quality of life (Euroquol-EQ-VAS).

The Pearson's correlation coefficients between the values of the outcome collected at baseline show there is a strong correlation between the three outcomes (BDI-II/WSAS $r = 0.753$; WSAS/EQ-VAS $r = -0.623$; BDI-II/EQ-VAS $r = -0.605$). There was missing data for each of the three outcomes. In total, 431 participants provided follow up data, however, there were only 429, 428 and 415 values recorded for BDI-II, WSAS and EuroQol respectively. 412 participants provided data for all three outcomes.

2.5.2 Ten Top Tips trial

The Ten Top Tips (10TT) is also a two-arm, individually randomised controlled trial (Beeken et al., 2012, Beeken et al., 2017). This dataset includes a sample of obese patients taken from 14 general practices across England. The general aim of the study was to investigate the effect of the 10TT intervention in primary care. The 10TT intervention consisted of a leaflet (called '*Ten Top Tips'*) listing target behaviours alongside advice on repetition and context stability.

The 10TT study specified weight change as the primary outcome. However, the trial team were interested in using three outcomes: change in weight (kg), change in waist circumference (cm) and blood glucose level (mmol/L). The weight and waist circumferences outcomes are viewed as continuous outcomes. The blood glucose level is categorised into 'standard' and 'high' groups (85% of the participants were categorised as standard). High blood glucose has been defined as levels greater than 7.0 mmol/L (WHO, website). The Pearson's correlation coefficients between the outcomes at baseline show a participant's weight is strongly correlated with a participant's waist circumference ($r = 0.775$). There is weak/moderate correlation between the participant's blood glucose level and a weight r$= 0.280$) and participant's blood glucose level and waist circumference ($r = 0.356$). The primary outcome was measured at three months. At this follow up, 388 participants provided at least one outcome value. However, only 383, 378 and 330 values were provided for weight, waist circumference and blood glucose level respectively.

2.6 Methods to analyse multiple outcomes in randomised controlled trials

In this section, relevant methods which can be used to analyse multiple correlated outcomes are described. A example of an illustrative trial is used to help describe the methods. In the illustrative two-arm trial there are m primary outcomes, which are correlated. The i^{th} trial participant is randomly assigned to either the intervention group $(x_i = 1)$ or the control group $(x_i = 0)$, for $i = 1, \dots, n$. Here, x_i is an indicator variable and n is the number of participants.

The aim of the trial is to test the null hypotheses $H_{0j}: \beta_{1j} = 0$ for $j = 1, \dots, m$ which state that there is no effect of the intervention on the nominated outcome. Each test statistic t_j is used to test the null hypothesis H_j. Further suppose that there is an overall null hypothesis $H(m) = \bigcap_{j=1}^{m} H_j$. Under this overall hypothesis, the joint test statistic $(t_1, \dots, t_m)$ has an M-variate distribution. Let Y_{ij} represent the outcome values corresponding to the i^{th} participant and the j^{th} outcome, β_{1j} represent the effect of the intervention on the j^{th} outcome and β_0 be the intercept term. Lastly, p_j is the j^{th} marginal unadjusted p-value which is obtained from the appropriate statistical test associated with analysing the respective outcome. To analyse a continuous outcome an unpaired Student's t-test may be used. To analyse a binary outcome a Chi-squared test may be used to investigate the intervention. The unadjusted statistical significance level is set to α. For simplicity, the subscript i associated with participants are omitted in most of the models and additional covariates have not been included in the models.

Unless otherwise stated, it is assumed that the intervention is shown to be effective if a statistically significant effect is found on at least one of the outcomes. Consequently, when referring to power for multiple comparisons disjunctive power is used, unless otherwise specified. These are the recommendations suggested by (Dmitrienko et al., 2009, Bretz et al., 2010).

2.6.1 Combine outcomes

One approach to avoid the difficulties associated with multiple significance testing is to combine the outcomes to create a single composite outcome. It avoids the issue of testing multiple outcomes as only one test is performed (Phillips and Haudiquet, 2003). A composite outcome is defined as the union of the outcomes. Consequently, if a composite outcome is made up of two time-to-event outcomes, then the composite outcome is defined as either

event occurring or both events occurring. However, for simplicity investigators may only take into account the time until the first event (Dmitrienko et al., 2009).

An example of a composite outcome is the time from randomisation until either a nonfatal ischemic stroke, fatal ischemic stroke or early death. Composite outcomes that combine several binary events, such as the previous example, most commonly arise in cardiovascular trials (Cordoba et al., 2010) or chronic disease trials (Cutter et al., 1999). When the composite outcome is the time until the first event there is an increase in statistical efficiency, compared to selecting only one of the events as the primary outcome. The increase in efficiency arises from the increased event rate. This may reduce the required sample size and consequently the costs and duration of the RCT (Ferreira-González et al., 2007, Freemantle et al., 2003).

The composite outcome needs to be specified before the trial begins andall components should be of equal importance when assessing the effect of the intervention (Montori et al., 2005). A composite outcome may not be appropriate when the effects of an intervention differ in magnitude and/or direction across the outcomes (Pogue et al., 2012). In particular, the latter may result in a large loss of power.

2.6.2 Analysing outcomes separately

It is common practice to analyse each outcome separately in a univariate framework. For example, when analysing continuous outcomes, an unpaired Student's t-test may be performed to analyse the effect of the intervention for each of the pre-specified outcomes. One typically would like to adjust for baseline covariates (European Medical Agency, 2017) in which case a classic linear model is preferable. For the scenario of continuous outcomes, the linear model can be written as

$$Y_j = \beta_{0j} + \beta_{1j}x + \epsilon_j \,, \tag{2.1}$$

where Y_j, x, β_0 and β_{1j} are as previously defined and $\epsilon_j \sim N(0, \sigma_e^2)$ is the random error with variance σ_e^2. By analysing the outcomes separately, the possible multivariate structure in the data has not be used. Indeed, an additional imputation step would be required to take into account any missing values.

Multiple tests need to be performed to analyse multiple outcomes in a univariate framework, which may increase the familywise error rate (FWER). A variety of techniques can be used to

ensure that the error rate is kept to an acceptable level. These techniques are discussed in detail later in this section.

2.6.3 Multivariate analysis

More advanced techniques, including multivariate methods (Goldstein, 2011), have been proposed that enable multiples outcomes to be analysed simultaneously by taking into account the correlations between them (Teixeira-Pinto et al., 2009). The use of these methods could potentially lead to improved precision and greater power (McCulloch, 2008b) and hence smaller sample sizes. In addition, depending on the objective of the trial, we may also estimate an overall effect of the intervention across outcomes, as well as a separate effect for each outcome.

Global statistical tests

Another approach is to use a global testing procedure to estimate an overall effect of the intervention across outcomes, with. the trial deemed a success if the overall effect is statistically significant. Conceptually,the interpretation of results obtained from global procedures and the analysis of composite outcomes are similar and both avoid the issues associated with testing outcomes separately. However, unlike composite outcomes, global test procedures account for the correlations between the outcomes. Methods include the multivariate analysis of variance (MANOVA), the one-degree of freedom global test developed by Roy (Roy et al., 2003), and the test statistics developed by O'Brien (O'Brien, 1984) and extended by Pocock (Pocock, 1997).

Global testing procedures require balanced data across all outcomes and will omit observations if any outcome values are missing. Given this limitation, global testing procedures are not widely used in clinical trials and therefore are not discussed further.

Multivariate regression

Multivariate regression is an extension of the multiple regression that allows for multiple outcomes of the same type to be analysed. For example, this approach may be used to analyse several continuous or several binary outcomes. To model the effects of the intervention for two continuous outcomes, the following model can be used

$$\boldsymbol{Y} = \boldsymbol{X_1}\boldsymbol{\beta}_0 + \boldsymbol{X_2}\,\boldsymbol{\beta_1} + \mathbf{E}, \tag{2.2}$$

where $\mathbf{Y}$ is a $n \times 2$ matrix in which each row contains the outcome values for a single participant (for two outcomes, $Y_i = (y_{i1}, y_{i2})$), $\boldsymbol{X_2}$ is a $n \times 1$ column vector in which each element indicates whether the participant received the control ($x_i = 0$) or intervention ($x_i = 1$), $\boldsymbol{\beta}_1$ is a $1 \times n$ row vector representing the effects of the intervention for the 2 outcomes, $\boldsymbol{X_1}$ is a $n \times 1$ column vector of 1s, $\boldsymbol{\beta}_0$ is a $1 \times n$ row vector representing the constant term (the 'intercept') for the 2 outcomes and $\mathbf{E}$ is a $n \times 2$ matrix representing the random error. The rows of the error term, $\mathbf{E}$, are independently distributed as multivariate normal $N_2(0, \boldsymbol{\Sigma})$ with an unknown positive definite covariance matrix $\boldsymbol{\Sigma}$.

Equation (2.2) can be adapted to handle multiple binary outcomes by assuming $\boldsymbol{Y}$ is a latent variable, such that the observed binary outcome $Y_{\boldsymbol{ij}}^* = 1$ if $Y_{ij} > 0$, and $Y_{\boldsymbol{ij}}^* = 0$ otherwise. Multivariate regression also requires balanced data across the outcomes.

Factorisation modelling

This approach involves factorisng the joint distribution of two correlated outcomes into a marginal and a conditional distribution. Univariate models can then be fitted to both components of the factorisation (Teixeira-Pinto and Harezlak, 2013). IT is possible to use different types of outcomes within this framework althoughthe estimated intervention effects are likely to be different from those obtained by modelling the outcomes separately because of different distributional assumptions. For the univariate analyses, it is assumed that the two outcomes are separate, whereas the factorisation model assumes that the second outcome is distributed conditionally on the first outcome (Teixeira-Pinto and Harezlak, 2013).

With two correlated outcomes, where one is continuous (Y_1) and the other is binary (Y_2) we can use one of the two possible factorisations of their joint distribution $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1 \mid Y_2}(y_1|y_2) f_{Y_2}(y_2)$. Fitzmaurice and Laird (1995) describe the factorisation model which uses a linear model for Y_1 and a probit model for Y_2, and including one covariate for the intervention. The model is

$$Y_1 = \beta_{01} + \beta_{11} x + \tau (Y_2 - \mu_2) + \epsilon_1, probit(\mu_2) = \beta_{02} + \beta_{12} x, \qquad (2.3)$$

where $\epsilon_1 \sim N(0, \sigma_c^2)$ is a normally-distributed random variable with mean zero and variance σ_c^2, and τ quantifies the association between Y_1 and Y_2 .

Catalano and Ryan (1992) propose the 'reverse' of this model in which they use the other possible factorisation of the joint distribution $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2 \mid Y_1}(y_2|y_1)$. This is

described in Teixeria-Pinto (2013). At present there is no guidance on how to analyse more than two outcomes using the factorisation model. With k outcomes, there are $k!$ possible factorisations and there is no guidance as to which one is best to use.

Latent variable model

Several researchers have suggested several methods that use latent variables to model multiple correlated outcomes, including Sammel et al. (1997), McCulloch (2008a) and Dunson (2000).

McCulloch (2008a) suggest introducing a random effect l that will be shared across outcomes. Assuming we have values of one continuous normally distributed outcome (Y_1) and a binary outcome (Y_2), the model is

$$\begin{aligned} &, \\ Y_1 &= \beta_{01} + \beta_{11}x + l + \epsilon_1, \\ &, \\ P(Y_2 = 1) &= \phi(\beta_{02} + \beta_{12}x + \lambda\, l), \end{aligned} \qquad (2.4)$$

where $e_1 \sim N(0, \sigma_1^2)$, $l \sim N(0, \sigma_l^2)$ and σ_1^2 and σ_l^2 are unknown variances. It is assumed that latent variable l completely specifies the pairwise correlation between the outcomes and hence, conditional on this variable, the two outcomes are independent. The parameter λ accounts for the fact that the linear predictors for Y_1 and Y_2 are on different scales and will therefore have different variances.

The estimated intervention effects for this model are conditional on the latent variable, as shown in equation (2.4) andconsequently, they may not be comparable to the estimates obtained from the other methods discussed. To obtain effects for the binary outcomes that are comparable to those obtained from univarate analyses, we divide the regression coefficient β_{12} by $\sqrt{\lambda^2\sigma_l^2 + \sigma_2^2}$ (Teixeira-Pinto and Normand, 2009). where σ_2^2 is fixed to 1 if a probit link function is used, or to $\frac{\pi^2}{3}$ if a logit link is used. A detailed discussion regarding the adjustments can be found in Teixeira-Pinto and Normand (2008).

Note that in the above example (2.4) there are four variance-covariance parameters: σ_1^2, σ_2^2, σ_l^2 and λ. In this example σ_2^2 is fixed due to the binary nature of the equation but the other three parameters need to be estimated. There are only two estimable quantities: the total residual of y_1

$$var(\epsilon_1) = \sigma_1^2 + \sigma_l^2,$$

and the correlation between the total residuals of the two equations

$$corr(\epsilon_1, \epsilon_2) = \frac{cov(\epsilon_1,\epsilon_2)}{\sqrt{var(\epsilon_1)var(\epsilon_2)}} = \frac{\lambda\sigma_l^2}{\sqrt{(\sigma_1^2+\sigma_l^2)(\sigma_2^2+\lambda^2\sigma_l^2)}}.$$

It is necessary to impose an additional constraint to ensure the model is not over parameterised so that the model parameters are identifiable (Teixeira-Pinto and Normand, 2009). One option is to fix the variance of the latent variable σ_l^2. A similar restriction is needed to analysing multiple continuous or multiple binary outcomes.

McCulloch (2008a) study this model in more depth and provide examples using other distributions. Sammel et al. (1997) discuss another latent variable model for mixed discrete and continuous outcomes. Their model allows use of any distribution from the exponential family.

Multivariate multilevel model

The multivariate multilevel (MM) model has been suggested as another approach to analyse correlated multiple outcomes. In the MM the multiple outcomes are considered to be nested within individuals and are treated in a similar manner to how repeated measurements are treated within the multilevel modelling framework(Goldstein et al., 2009, Goldstein, 2011).

For two continuous outcomes, the following model is used

$$Y_j = z_{1j}(\beta_{01} + \beta_{11}x + \epsilon_1) + z_{2j}(\beta_{02} + \beta_{12}x + \epsilon_2), \qquad (2.5)$$

$$z_{1j} = 1 \text{ if } j = 1 \text{ and } z_{1j} = 0 \text{ otherwise,}$$

$$z_{2j} = 1 - z_{1j},$$

where $j \in \{1,2\}$ indicates the outcomes, z_{kj} is an indicator for outcome Y_j x_i is the binary variable indicating whether the participant received the control ($x_i = 0$) or intervention ($x_i = 1$), β_{1j} is the effect of the intervention and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega_u})$ is the random error for the level 2 structure where $\boldsymbol{\Omega_u}$ is the unknown covariance matrix.

Level one variation is not specified as the level exists solely to define the multivariate structure. The formulation as a multilevel model allows for estimation of a covariance matrix even if some of the outcome data are missing, as long as missing at random. In the above model, two intervention effects have been specified, one for each outcome. However, a common effect across both outcomes may also be specified. Additionally, the model can be

extended to incorporate multiple covariates (Rasbash et al., 2012) and the model can handle mixed outcome types (Goldstein et al., 2009).

Summary of the multivariate methods

The factorisation, latent and MM models can handle continuous outcomes, binary outcomes or a mixture of both. In addition, these models can handle non-overlapping missingness, where values may be missing for some but not all some of the outcomes. That is, the number of observations does not need to be balanced across outcomes. The multivariate regression, latent and MM models can easily be extended to several outcomes although the factorization can be cumbersome when there are several outcomes. As the factorization model cannot be extended to more than two outcomes, I encourage the use of the latent or MM model.

When the effects are analysed separately an adjustment, such as those described below in Section 2.7, will need to be made for multiple comparisons to control FWER.

2.7 Methods to control the familywise error rate

In the previous section, I discussed methods to analyse multiple outcomes in the trial setting. When implementing many of these methods an intervention effect is estimated on each outcome and multiple comparison are performed. In confirmatory RCTs, the FWER must be maintained at an acceptable level which is usually 0.05. In this section, methods that may be used to control the FWER are described.

2.7.1 Hierarchical testing of multiple outcomes

Hierarchical testing involves ranking the outcomes according to their clinical relevance. The outcomes are ranked from most important to least important and then tested individually in the pre-specified hierarchical order. An outcome can only be tested if all previously tested outcomes have been shown to be statistically significant; otherwise the testing stops and no confirmatory claims can be based on the remaining outcomes. For example, if Y_1 and Y_2 are ordered to reflect clinical importance, the intervention effect on Y_2 can only examined if the intervention effect on Y_1 was found to be statistically significant. Because of the hierarchical nature, the same significance level can be used for all tests and no formal adjustment is necessary. However, the power is reduced for outcomes that have lower ranks. Other methods have been introduced to maintain some of the power for the outcomes with lower

ranks, including the fixed-sequence and fallback methods (Dmitrienko and D'Agostino, 2013).

In some instances, there is a natural hierarchical order for the outcomes, for example, with respect to the clinical importance of the outcomes, and therefore it is suitable to use this technique to remove the problem of multiple comparisons. If investigators use this method, the ordering must be pre-specified and clearly reported in the study protocol (Committee for Proprietary Medicinal Products, 2002).

2.7.2 Adjustment to the p-values

To take account of multiple comparisons, the FWER may be controlled by applying a method to adjust the p-values produced by each statistical test used to investigate the effect of the intervention or equivalently the corresponding significance levels may be adjusted. Many techniques to adjust p-values have been proposed in the literature (Dmitrienko and D'Agostino, 2013, Shaffer, 1995, Dmitrienko et al., 2009). The techniques can be categorised into single step methods that test all hypotheses simultaneously, stepwise methods that rely on data-driven hypothesis ordering and stepwise methods that rely on a pre-specified hypothesis ordering. Once the p-values have been adjusted for multiplicity, the intervention can be deemed to be effective if a statistically significant effect if found on at least one of the outcomes. The single step procedures are described below.

Šidák method

The Šidák method (Šidák, 1967) is a single step adjustment method. The adjusted p-value is given by

$$p_j^{\check{S}i} = 1 - \left(1 - p_j\right)^m, \tag{2.6}$$

where $p_j^{\check{S}i}$ is the Šidák p-value adjusted for multiplicity. The p-value $p_j^{\check{S}i}$ should be compared to the nominal significance level. Equivalently, the significance level could be adjusted so that the unadjusted p-values are compared to the adjusted significance level

$$\alpha_j^{\check{S}i} = 1 - \left(1 - \alpha_j\right)^{\frac{1}{m}}. \tag{2.7}$$

Under the assumption that the outcomes are independent, the method can be derived, as follows:

$$P(\text{no Type I error on 1 test}) = 1 - \alpha_j^{\check{S}i}, \tag{2.8}$$

$$\rightarrow P(\text{no Type I error on m tests}) = \left(1 - \alpha_j^{\check{S}i}\right)^m,$$

$$\rightarrow P(\text{at least one type I error on m tests})$$
$$= 1 - \left(1 - \alpha_j^{\check{S}i}\right)^m = \alpha_j.$$

The final line in equation (2.8) gives the result given in equation (2.7). By using the smaller significance level $\alpha_j^{\check{S}i}$ for each outcome, the overall significance level is maintained at the nominal level. Figure 2.2 provides a graphical summary of the Šidák method when there are two outcomes. The Šidák equation was derived under the assumption that the outcomes are independent but also controls the FWER when the hypothesis test statistics are multivariate normal.

Bonferroni method

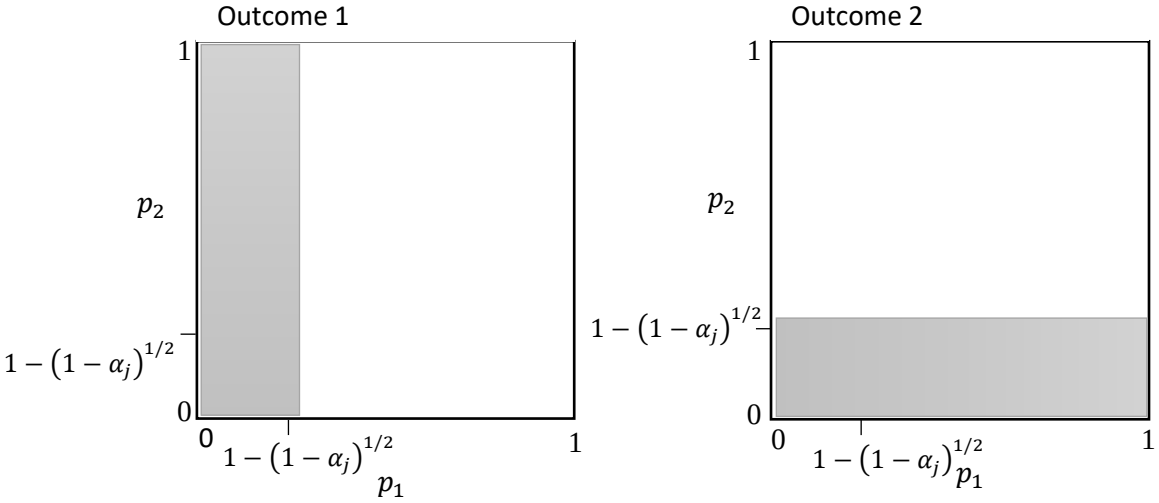
The most basic single step adjustment method is the Bonferroni method. It relies on a simple α splitting rejection rule. The adjusted p-value is defined as

$$p_j^{Bonf} = m\, p_j. \tag{2.9}$$

The adjusted p-value p_j^{Bonf} should be compared to the pre-specified significance level. Equivalently, the significance level could be adjusted so that the unadjusted p-values are

Figure 2.2 Graphical summary of the Šidák method for two outcomes.

The shaded area is the combination of p-values for which the null hypothesis is rejected, which is there is no effect of the intervention for the corresponding outcome. The rejection region for outcome 1 is displayed on the left and the corresponding rejection region for outcome 2 is displayed on the right. Similar graphs are displayed in Dmitrienko and D'Agostino (2013).



compared to the adjusted significance level, for example, with two outcomes we would use an adjusted significance level $\alpha^{Bonf} = \frac{\alpha_j}{2} = \frac{0.05}{2} = 0.025$ where α_j is the unadjusted significance level. The method can be derived using the Taylor series expansion from the Šidák equation, as

$$\begin{aligned} \alpha_j^{\check{S}i} &= 1 - \left(1 - \alpha_j\right)^{\frac{1}{m}} \\ &= 1 - \left(1 + \left(-\alpha_j\right)\right)^{\frac{1}{m}} \\ &= 1 - \left(1 + \frac{1}{m}\left(-\alpha_j\right) + \left(\frac{1}{m}\right)\left(\frac{1}{m} - 1\right)\left(-\alpha_j\right)^2 + \cdots\right) \\ &\approx \frac{\alpha_j}{m} \\ \alpha_j^{Bonf} &= \frac{\alpha_j}{m}. \end{aligned} \tag{2.10}$$

Figure 2.3 Graphical summary of the Bonferroni method for two outcomes.

The shaded area is the combination of p-values for which the null hypothesis is rejected, that is there is no intervention effect for the corresponding outcome. The rejection region for outcome 1 is displayed on the left and the corresponding rejection region for outcome 2 is displayed on the right. Similar graphs are displayed in Dmitrienko and D'Agostino (2013).

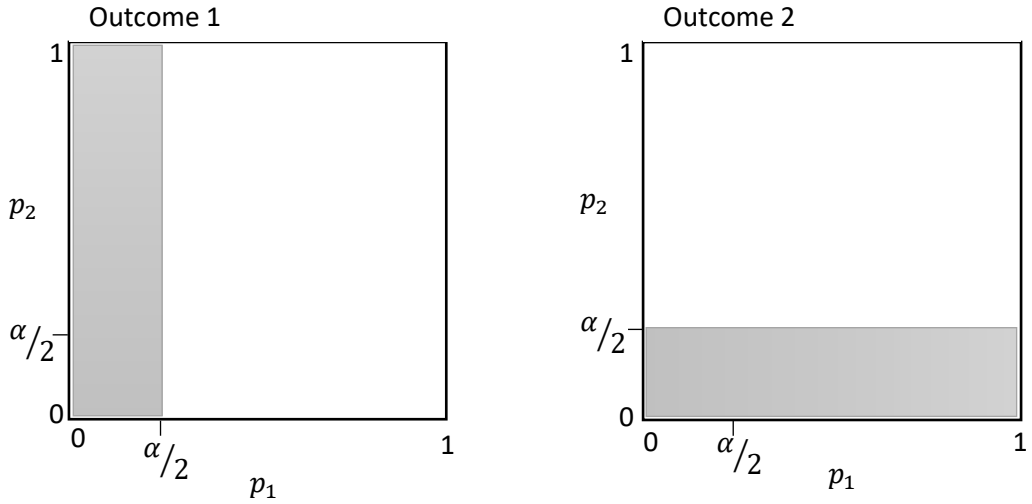


Figure 2.3 provides a graphical summary of the Bonferroni method when there are two outcomes. The advantage of this method is that it is simple and it is a non-parametric method. As it a non-parametric method, it does not impose any restrictions of the type of test required or distribution of the test statistics. Given its simplicity, the Bonferroni method

is widely used in RCTs even though it can be conservative, when the outcomes are correlated or when the number of tests is large (Yoon et al., 2011, Tyler et al., 2011).

The Bonferroni method is less powerful than the Šidák method since

$$\alpha_j^{Bonf} = \frac{\alpha_j}{m} < 1 - \left(1 - \alpha_j\right)^{\frac{1}{m}} = \alpha_j^{\check{S}i}, \tag{2.11}$$

as shown in the Taylor expansion in equation (2.10). However, the improvement compared to the Bonferroni method is minimal, especially when there are less than ten tests (Simes, 1986).

Derivatives of the Šidák method

The Dubey/Armitage-Parmar (D/AP), Tukey, Ciminera, Heyse (TCH), (Tukey et al., 1985); and the R2-adjustment (RSA), (Sankoh et al., 1997) adjustments are *ad-hoc* methods that are based on the Šidák method, which takes into account the correlation between the outcomes. These methods have the form

$$p_j^{adj} = 1 - \left(1 - p_j\right)^{g(j)}. \tag{2.12}$$

where $g(j)$ is defined for each method. The D/AP method defines $g(j)$ as $m^{1-\mu_j}$ where μ_j is the mean correlation between the j^{th} outcome and the remaining $m-1$ outcomes. When using this method in the analysis of multiple outcomes, the mean correlation may be estimated from the data to calculate the adjusted p-values. The TCH has been derived for outcomes that have a strong correlation, it defines $g(j)$ as $\sqrt{m}$ where m is the number of outcomes. The RSA defines $g(j) = m^{1-R2(j)}$ where $R2(j)$ is the value of R^2 from an intercept-free linear regression with the j^{th} variable as the outcome and the remaining $m-1$ variables as the predictors.

There has been little theoretical work to assess the performance of these methods (Sankoh et al., 1997). One of the nice properties of the D/AP procedure, which may have contributed to its development, is that when the mean correlation between the j^{th} outcome and the remaining $m-1$ outcomes is 0, then the D/AP method is equivalent to the Šidák method. On the other hand, when the mean correlation between the j^{th} outcome and the remaining $m-1$ outcomes is 1, then the D/AP is reduced to $p_j^{adj} = p_j$, meaning that the adjusted p-value is the same as the p-value unadjusted for multiplicity. When the mean correlation between the j^{th} outcome and the remaining $m-1$ outcomes is 0.5 this method coincides with the TCH method.

The methods described so far have been single step methods. The data-driven methods that require the data to be ordered before implementing the adjustment are now described.

Holm method

The Holm 'step-down' method (Holm, 1979) is a data-driven stepwise method that is also known as the 'sequentially rejective Bonferroni' test.

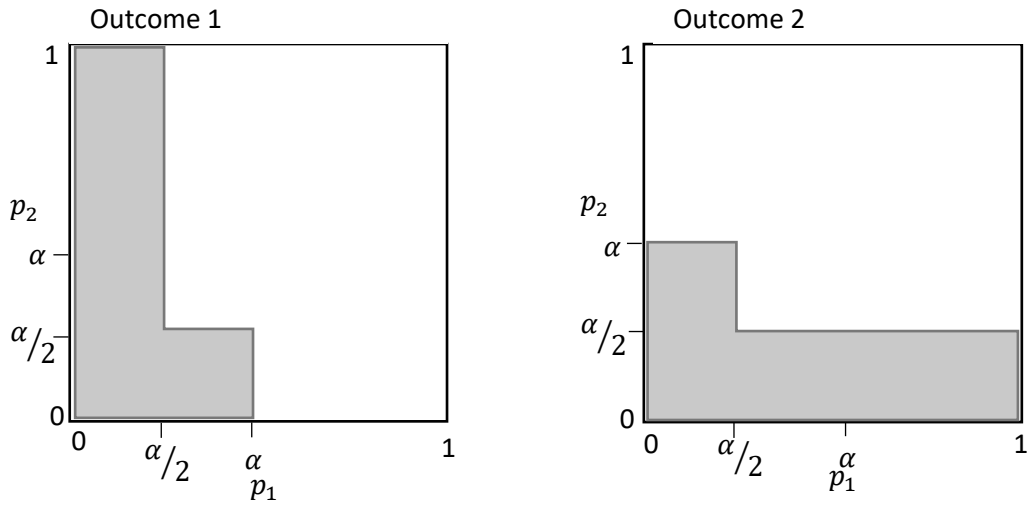
For this method, the p-values unadjusted for multiplicity are ranked from smallest $p_{(1)}$ to largest $p_{(m)}$ and adjusted as follows

$$p_{(k)}^{Holm} = (M - k + 1)\, p_{(k)}, \tag{2.13}$$

where $p_{(k)}$ is the unadjusted p-values corresponding to the outcome value $Y_{(k)}$ for $k = 1, \ldots m$, the rank of the p-value, and m is the number of outcomes. Starting with the most significant p-value (smallest p-value), each p-value adjusted for multiplicity is compared to the pre-specified significance level, until a p-value greater than the significance level is observed after which the procedure stops (Wright, 1992). The Holm method is described graphically in Figure 2.4.

Figure 2.4 Graphical summary of the Holm method for two outcomes.

The shaded area is the combination of p-values for which the null hypothesis is rejected, that is there is no effect of the intervention for the corresponding outcome. The rejection region for outcome 1 is displayed on the left and the corresponding rejection region for outcome 2 is displayed on the right.



As with the Bonferroni method, the Holm method is a non-parametric method and therefore does not impose any restrictions on the distribution of the joint test statistic.

The Holm method is more powerful than the simple Bonferroni method (Yoon et al., 2011) meaning that if a null hypothesis is rejected when using the Bonferroni method, the null hypothesis will also be rejected by the Holm method but additional hypothesis may be rejected when using the Holm method. This is shown graphically, as the shaded region is larger in Figure 2.4 which shows the rejection region for the Holm method compared to Figure 2.3 which shows the rejection region for the Bonferroni method.

Hochberg Step-Up method

The Hochberg step-up method (Hochberg, 1988) is analogous to the Holm step-down method. For this method, the p-values unadjusted for multiplicity are ranked from largest $p_{(1)}$ to smallest $p_{(m)}$ and adjusted as follows

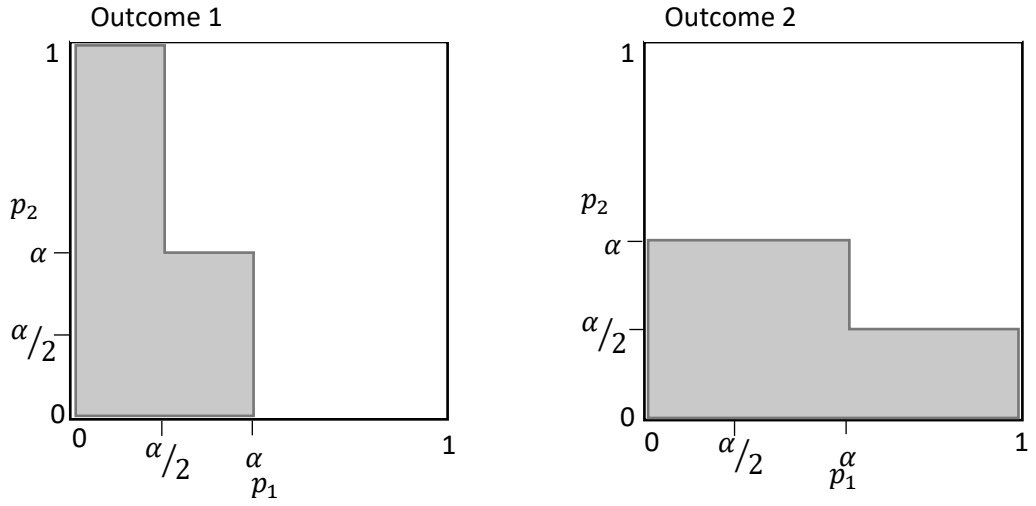
$$p_{(k)}^{Hoch} = (m - k + 1)\, p_{(k)} \tag{2.14}$$

where $p_{(k)}$ is the unadjusted p-values corresponding to the outcome value $Y_{(k)}$ for $k = 1, \ldots, m$ which is the rank of the p-value. Starting with the least significant p-value (largest p-value), each p-value adjusted for multiplicity is compared to the pre-specified significance level, until a p-value *lower* than the significance level is observed after which the comparison

stops (Wright, 1992). Once the procedure stops, the current outcome is defined as statistically significant and all remaining outcomes (which have a p-value less than or equal to the one being tested) are defined as statistically significant. This is a semi-parametric method meaning it can be used to control the FWER when the distribution of the joint test statistic under the alternative hypothesis is known (for example multivariate normality) but not fully specified. If the Hochberg method is applied to multiplicity problems with a negatively correlated test statistic, the FWER may be inflated, however, the magnitude of the error rate inflations with negative correlation is typically trivial (Dmitrienko and D'Agostino, 2013). The Hochberg method is described graphically in Figure 2.5.

Figure 2.5 Graphical summary of the Hochberg method for two outcomes.

The shaded area is the combination of p-values for which the null hypothesis is rejected, that there is no effect of the intervention for the corresponding outcome. The rejection region for outcome 1 is displayed on the left and the corresponding rejection region for outcome 2 is displayed on the right.



The Hochberg method is more powerful than the Holm method (Candes, 2012). This is highlighted by a larger shaded region in Figure 2.5 compared to Figure 2.4. This means that when using the Hochberg method one is guaranteed to reject all null hypotheses that are rejected when using the Holm method, but additional null hypotheses may also be rejected when using the Hochberg method.

The Hochberg method favours consistency among the outcomes across multiple tests in the sense that it is easier to achieve significance if all p-values are small. Whenever all p-values in a multiplicty problem are significant before an adjustment (i.e. none of the p-values

exceed α), all hypothesis will be rejected after the Hochberg method is applied (Dmitrienko and D'Agostino, 2013).

Hommel method

The Hommel method (Hommel, 1988) is another data-driven stepwise method. For this method, the unadjusted p-values are ranked from largest $p_{(m)}$ to smallest $p_{(1)}$. Let l be the largest integer for which

$$p_{(m-l+j)} > \frac{j\alpha}{l}, \tag{2.15}$$

for $j = 1, \dots l$. If no such j exists then all outcomes can be deemed statistically significant; otherwise, all outcomes with $p_i \leq \frac{\alpha}{j}$ may be deemed statistically significant, for $j = 1, \dots, m$ and $i = 1, \dots, m$.

The Hommel method has greater power to detect a true effect of the intervention compared to the Hochberg method (Dmitrienko and D'Agostino, 2013). Similarly to the Hochberg method, it is a semi-parametric method meaning it can be used when the distribution of the test statistic under the alternative hypothesis is known but not fully specified. Additionally, it requires consistency among the outcomes of the individual tests as it is easier to achieve significance if the p-values for all hypotheses to be performed are small (Dmitrienko and D'Agostino, 2013).

Another class of methods to account for multiple comparisons is the resampling method. The resampling methods take into account the correlation between the outcomes via bootstrapping (Westfall and Young, 1993). I will consider one resampling method below.

Stepdown MinP

Another step-down method to adjust p-values is the 'stepdown MinP' method (Westfall and Young, 1993, Ge et al., 2003). Unlike the previous methods, it does not make any assumptions regarding the joint distribution of the test statistics, instead it attempts to approximate the true joint distribution by using a resampling approach. Consequently, the stepdown MinP is referred to as a 'resampling based procedure' (Dmitrienko et al., 2009). The resampling based procedure takes into account the correlation structure between the outcomes and therefore may yield more powerful tests compared to the other adjustment methods (Reitmeir and Wassmer, 1999). The steps to obtain the stepdown MinP p-value adjusted for multiplicity are: 1) calculate the observed test statistics for the observed

dataset; 2) resample the data with replacement within each group to obtain bootstrap resamples, compute the resampled test statistics for each resampled dataset and construct the reference distribution using the centred and/or scaled resampled test statistics; 3) calculate the critical value of a level α test based on the upper α percentile of the reference distribution, or obtain the raw p-values by computing the proportion of bootstrapped test statistics that are as extreme or more extreme than the observed test statistic (Li and Dye, 2013).

The resampling techniques have previously been recommended for clinical trials with multiple outcomes (Reitmeir and Wassmer, 1999) however, they are not widely used in clinical trial applications. The stepdown MinP was the only resample method discussed as it has been shown to perform well when compared to other resampling methods (Li and Dye, 2013).

All the methods that have been discussed so far have assumed that the study has been powered adequately for all primary outcomes. However, this may not be the case. In some scenarios, one outcome may have adequate power whilst the remaining primary outcomes are underpowered due to time and cost restraints. Alternatively, the study may not have been powered to investigate secondary outcomes but the investigator is still interested in exploring the effects of the intervention on the secondary outcomes. Prospective alpha allocation scheme and the adaptive alpha allocation approach are designed to be used in the scenario when some outcomes are underpowered.

Prospective alpha allocation scheme (PAAS)

The prospective alpha allocation scheme (PAAS) is a weighted version of the Bonferroni method (Moyé, 2000). For this approach, the outcomes have to be ranked in order of priority, with the most important outcome being ranked first. For two outcomes the approach is defined as:

i) α_1 is chosen as the significance level for the most important outcome, where $0 < \alpha_1 < \alpha$,

ii) the second outcome has the following significance level:

$$\alpha_2 = 1 - \frac{1-\alpha}{1-\alpha_1},$$

where α is the pre-specified level of FWER which is usually 0.05. An extension for a larger number of primary and secondary outcomes is provided in Moyé (2000). This approach is

useful when the outcomes can be hierarchically ordered in order of importance, for example, when there is one primary outcome and a key secondary outcome. However, it has limited use when alternative outcomes are used as the investigators may find it difficult to hierarchically order the outcomes or chose the level of α to give each outcome. The PAAS is a simple way to accommodate outcomes that can be ordered a priori but it potentially under powers outcomes (Li and Mehrotra, 2008), consequently other approaches, including the adaptive alpha allocation approach have been proposed.

Adaptive alpha allocation approach (4A)

The adaptive alpha allocation approach (4A) is a feedback procedure (Li and Mehrotra, 2008). The method assumes that the outcomes can be grouped into two families. The first family includes primary outcomes that are adequately powered and the second family includes potentially underpowered outcomes, potentially the secondary outcomes. For two outcomes, the p-values unadjusted for multiplicity are ranked $p_{(1)}$ and $p_{(2)}$ according to the importance of the corresponding outcome. Assuming the outcomes are independent, the approach is defined as:

i) The most important outcome is tested using $\alpha_1 = \alpha - \epsilon, \ \epsilon > 0$

ii) The least important outcome is tested using

$$\alpha_2 = \begin{cases} \alpha & \text{if } p_{(1)} \leq \alpha_1 \\ \min\left(\frac{\alpha_t}{p_1^2}, \alpha_1\right) & \text{if } p_{(1)} > \alpha_1 \end{cases},$$

where

$$\alpha_t = \begin{cases} \alpha_1 \left(1 - \sqrt{2\,\alpha_1 - \alpha - \alpha_1^2}\right)^2 & \text{if } \alpha_1 + \alpha_1^2 - \alpha_1^3 \leq \alpha \\ \alpha_1 \frac{\alpha - \alpha_1}{1 - \alpha_1} & \text{if } \alpha_1 + \alpha_1^2 - \alpha_1^3 > \alpha \end{cases},$$

where α_i is the significance level corresponding to the outcome ranked i^{th} and α is the chosen FWER. Li and Mehrotra (2008) provide an extended version that takes into account any correlation between the outcomes. They provide tables describing the level of alpha to use depending on the correlation between the outcomes. This approach is beneficial as it provides higher significance level for the less important outcomes. However, this work focuses on multiple primary outcomes which often cannot be ordered according to their level of importance.

Other methods

There have been other methods that have been described in the literature to adjust p-values including Rom's test which is more powerful than the Hochberg, Holm and Bonferroni methods; however, it is more complicated than these methods (Wright, 1992). Dunnett (1955) has suggested a family of methods which require the outcomes to be normally distributed. The limitation of the Dunnett methods is that they require a balanced design in that they require the same number of observations for all the outcomes. In the clinical trial setting, this is often not achieved with missing variables arising for numerous reasons, consequently the Dunnett methods are not described in more detail.

Summary of methods which adjust p-values

In this section, I have described methods to account for multiplicity based on adjusting univariate p-values. When calculating the p-values, it has been assumed that the main clinical objective of the trial is formulated in terms of investigating the effect of the intervention on several primary outcomes and the objective is met if at least one analysis produces a significant result (Dmitrienko and D'Agostino, 2013). A summary of adjustments, including those described above, are shown in Table 2.1.

The Bonferroni and Holm methods are non-parametric. This means they both control the FWER in any setting. Hochberg and Hommel methods are semi-parametric and therefore certain distributional assumptions need to be satisfied to achieve the FWER control. Semi-parametric methods can be used when the distribution of the joint test statistic used to test the null hypothesis is known but it is not fully specified. For example semi-parametric methods can be used when it is known that the joint distribution of the test statistic is multivariate normal but the mean of this distribution is not known. These distributional assumptions are not restrictive and many clinical trials meet these assumptions (Dmitrienko and D'Agostino, 2013). Parametric methods can be used when the joint distribution of the test statistic is fully specified.

Table 2.1 Methods that can be used to control the familywise error rate (FWER) when analysing multiple outcomes in clinical trials

	Classification		
Distributional information	Single step	Data-driven hypothesis ordering	Pre-specified hypothesis ordering
Non-parametric	Bonferroni	Holm Stepdown MinP*	Fixed-sequence Fallback Chain
Semi-parametric	Šidák TCH* D/AP* RSA*	Hochberg Hommel	
Parametric	Dunnett	Step-down Dunnett Step-up Dunnett	Parametric fallback Parametric chain Feedback

This table is similar to a table from (Dmitrienko and D'Agostino, 2013) TCH = Tukey, Ciminera, Heyse; D/AP = Dubey/Armitage-Parmar; RSA = R2-adjustment. *These methods account for the correlations between outcomes.

Single step methods are inefficient because they do not utilize the α propagation and thus do not use up all of the available error rate. The stepwise methods are more powerful methods. As demonstrated by Figures 2.2-2.5, the common methods can be ordered in terms of increasing power: Bonferroni, Holm, Hochberg and Hommel. This highlights that Hommel and Hochberg are preferred over the other two methods in a multiplicity problem without hypothesis ordering. However these do require additional distributional assumptions over the other Bonferroni type methods (Dmitrienko and D'Agostino, 2013).

Fixed sequence methods are used when the outcomes are ordered in terms of importance prior to the trial. These methods are not as useful when alternative or co-primary outcomes are used as investigators are unlikely to be able to order the importance of the outcomes. Consequently, the fixed sequence methods are not discussed in detail.

When selecting the method to use the extent of the pairwise correlation between the primary outcomes needs to be considered along with the impact of the FWER. Most of the methods described above ignore the pairwise correlations. Ignoring these correlations could

result in a loss of efficiency and consequently less power being required to detect effects of the intervention (Teixeira-Pinto et al., 2009). The TCH method was designed for strongly correlated outcomes, consequently, when outcomes are independent the FWER observed is very high, nearly double the desired threshold (Blakesley et al., 2009). Researchers must be confident that the outcomes will be correlated if this method is chosen.

2.8 Discussion

Many techniques have been proposed to analyse multiple primary outcomes in clinical trials. The multivariate methods are more efficient compared to analysing the outcomes separately. The gains in efficiency may lead to smaller standard errors and, as a result, higher power. This in turn may affect the conclusions drawn. When analysing multiple outcomes in confirmatory randomised controlled trials, it is vital to control the FWER, to ensure that the chance of observing at least one statistically significant result by chance is not too high. Many p-value adjustment methods have been proposed to maintain the FWER. Even though many multivariate methods and methods to adjust p-values have been proposed, it is not known which of these methods, if any, are used in published clinical trials.

Chapter 3

A review of recently published randomised controlled trials

This chapter reviews recently published RCTs in the areas of neurology and psychiatry. The purpose of this chapter is to ascertain whether multiple primary outcomes are used in recently published trials, and to identify the methods that are used for the sample size calculation and for the statistical analysis of RCTs using multiple primary outcomes. In doing so, I will be able to ensure that the simulations in the following chapters are based on realistic scenarios to ensure their relevance. This will enable me to provide practical guidance for researchers that is applicable to current practice.

The review focuses on major neurology and psychiatry journals. Neurology and psychiatry are two disease areas where multiple outcomes may provide a more comprehensive understanding of the potential effects of the intervention (Blakesley et al., 2009, Teixeira-Pinto et al., 2009). More specifically, multiple outcomes may be beneficial in trials investigating interventions for depression (Tyler et al., 2011), stroke (Mayo and Scott, 2011) or long term mental health conditions (De Los Reyes et al., 2011). In these disease areas, multiple primary outcomes may be required to provide a comprehensive understanding of the effects of an intervention.

I am primarily interested in the sample size calculation and the statistical analysis used in recently published trials. As discussed in chapter 2, when multiple outcomes are used, it is essential that all primary outcomes are taken into account during the design and analysis of the trial. If all outcomes are not considered, then the chosen analysis may be inefficient or the error rates may be unacceptably high. The sample size is an important consideration during the design of a trial. A good choice of sample size in necessary to ensure that the trial is efficient, ethical and cost effective (Röhrig et al., 2010). The number of primary outcomes and their pairwise correlations should be considered when determining the sample size.

The work in this chapter has been published in Contemporary Clinical Trials (Vickerstaff et al., 2015). The full paper is provided in Appendix 1.

3.1 Methods

3.1.1 Selecting the journals

A number of journals were selected for having a high impact and for frequently publishing randomised trials in the fields of psychiatry and neurology. The impact factor used was based upon the Thomson Reuters Journal Citation Report, published in 2010. This report was selected as 2010 was the year before the articles included in the review were published (2011-2014). Thomson Reuters is the source of the annual Journal Impact Factors. By choosing journals with high impact factors, I am choosing high quality literature that is likely to be cited and used for further research.

The areas of neurology and psychiatry were selected as RCTs are common in these areas (Wittchen et al., 2011). Multiple outcomes are particularly common in these areas as one outcome is rarely able to satisfactorily describe the health condition being investigated (Blakesley et al., 2009).

After having reviewed the impact factors, the following journals were selected for the final review:

Psychiatry Journals:

1. The American Journal Psychiatry (Am. J Psych)
2. JAMA Psychiatry (JAMA Psych)

Neurology Journals:

1. The Lancet Neurology
2. Neurology

General medicine journals:

1. The New England Journal of Medicine (NEJM)
2. The Lancet

3.1.2 Search criteria

The journals were hand searched for reports of randomised trials published between July 2011 and June 2014 inclusive. The time frame was selected to ensure that the results represented the methods currently adopted in the literature, at the time of writing.

The review of journals included additional supplementary material, such as protocols and appendices provided that they were referred to in the paper. The following trials were excluded from the analyses: proof of principle trials; phase II trials, including pilot trials and small crossover trials; secondary analyses of trials. They were excluded as they are exploratory trials that lead on to confirmatory trials and they often have limited power. A study was classified as a pilot if it was clearly defined as such, or if it was described as an exploratory study prior to a larger study within the discussion section.

3.1.3 Outcomes

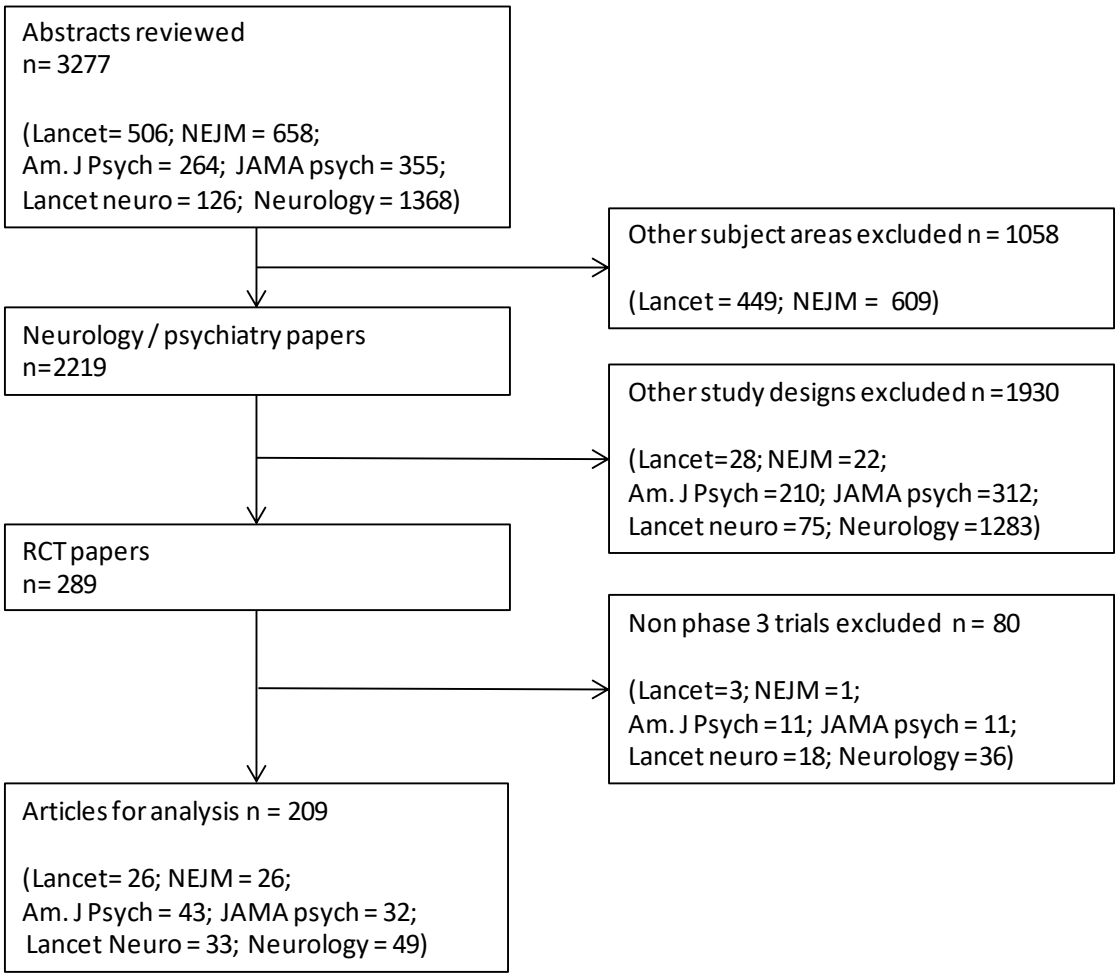
For each published trial, I examined the results in the abstract and the main text and the methods used for sample size calculation and statistical analysis. I recorded the number of primary and secondary outcomes and the methods used to account for multiple primary outcomes. Each outcome was recognised as primary if it was explicitly described this way or implicitly described this way by the aims of the trial. Otherwise, it was assumed that each outcome was primary. In the event that the primary outcomes described in the abstract differed to the main text, the outcomes reported in the main text were used.

I performed the initial assessments. For the trials where the primary outcomes were not clearly specified, the trials were appraised independently by other assessors (my supervisors Rumana Z. Omar and Gareth Ambler). All discrepancies were resolved by discussion between assessors. The statistical analyses were performed using Stata version 12 (StataCorp StataCorp).

3.2 Results

From the six journals, I reviewed a total of 3277 abstracts and identified 209 RCTs that met the inclusion criteria. Details of the study screening process can be seen in Figure 3.1.

Figure 3.1 Flow diagram of the screening process for the review of published RCTs.

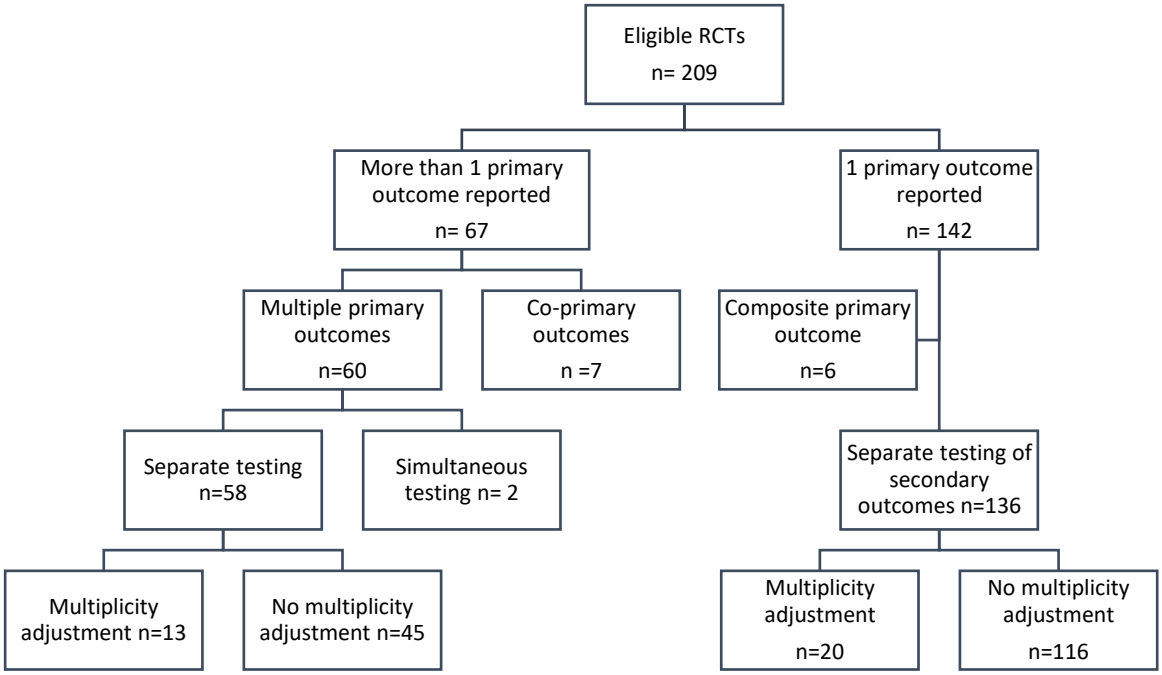


The majority of the trials (92%) were parallel-group, individually randomised trials, with a median number of subjects of 242 (IQR 112-549) and a median follow up time of six months (IQR 3-17.5 months); Table 3.1 and Figure 3.2 summarise the characteristics of these trials. A list of included studies can be found in Appendix 2.

Table 3.1 Description of the trials included in the review of published randomised controlled trials

Characteristic		Number (%) N = 209	
Journals	The New England Journal of medicine	26	(12)
	The Lancet	26	(12)
	The American Journal of Psychiatry	43	(21)
	JAMA psychiatry	32	(15)
	The Lancet neurology	33	(16)
	Neurology	49	(24)
Arms per trial	2	144	(68)
	3	52	(25)
	4+	15	(7)
Sites per trial	Single centre trial	36	(17)
	Multi-centre trial	173	(83)
Design of trial	Individually randomised, parallel design	193	(92)
	Individually randomised, factorial design	4	(2)
	Cluster randomised	12	(6)
Number of primary outcomes	1	142	(68)
	2	43	(21)
	3	14	(7)
	4	4	(2)
	≥ 5	6	(3)

Figure 3.2 Flow chart showing how the outcomes were analysed in the RCTs.



The trials included in the review were categorised by the number of primary outcomes (one or more than one), type of primary outcomes (alternative outcomes or co-primary outcomes), disease area of journal (psychiatry, neurology or general medicine) and intervention type (drug or non-drug). The following sections describe the results for each of these categories in turn.

3.2.1 Trials with no stated primary outcome or with multiple primary outcomes

Six of the 209 trials (3%) did not clearly specify a primary outcome. These trials did not follow the International Standards for Clinical Trials Registries produced by the World Health Organisation which states that both the primary and secondary outcomes should be defined and pre-specified (WHO, 2012). It was assumed that all the outcomes in these trials were equally important, and therefore recognised as primary outcomes.

Nearly a third of the examined trials (n=60, 29%) reported results for multiple primary outcomes. Forty-five (75%) of these 60 trials did not include adjustments for multiple comparisons. If multiple comparisons had been accounted for using the Bonferroni method, 6 of the 26 trials that reported that the intervention was effective would have drawn different conclusions. The results for one of these trials is described as a case study below. The remaining 15 (25%) trials accounted for multiple comparisons: six used the Bonferroni

method, seven used other adjustment methods (Holm, Hochberg-Benjamini, Šidák, Dunnett and sequential adjustments), and two performed MANOVA.

Some investigators stated that they did not adjust for multiple comparisons (Weiss et al., 2011, Nobile-Orazio et al., 2012, Dodel et al., 2013, Tariot et al., 2011, Gray et al., 2012). This suggests that some of the authors are aware the analysis may be different if they had adjusted for multiple outcomes. For example, Grey et al., (2012, p.5) wrote that "no adjustments for multiple testing were made, as they are known to reduce statistical power and increase the probability of accepting a null hypothesis that is truly false. Preliminary analyses leading to a priori hypotheses suggest that differences noted are less likely to be from chance alone". In this paper by Gray et al. (2012) adjustments were not made, presumably to achieve a statistically significant result even though the investigators were aware of the limitations of their analysis. Another justification provided for not accounting for multiple comparisons was "to prevent Type II error"(Vitiello et al., 2014).

The problem of multiplicity can be overcome by specifying different primary outcomes for different health features. For example, Launer et al. (2011) specified primary outcomes for cognitive measures and brain structure measures, respectively.

The abstracts of the trials were also examined to see if the investigators had specified that all the multiple outcomes were primary. The abstract summarises the paper and is often read in isolation from the main text. As such, the main outcomes should be clearly stated in the abstract (Hopewell et al., 2008). Just over half (57%, n=34) of the trials were found to clearly specify multiple primary outcomes in the abstract. The remaining abstracts described the outcomes, without specifying the order of importance, even though there were later specified as primary and secondary in the body of the papers.

In addition, the sample size calculations were reviewed. Fourteen (23%) of the 60 trials that reported multiple primary outcomes incorporated only one outcome in the sample size calculation. Fourteen of the trials clearly reported sample size calculations that incorporated more than one outcome. The methods to account for multiplicity in the sample size calculation included: a multiplicity-adjusted significance level in the calculation (Nierenberg et al., 2013); using simulations developed by Heo and Leon (2008), (Conrod et al., 2013); and calculating the sample size separately for each of the primary outcomes then selecting the largest value as the final sample size (Odekerken et al., 2012). Lovera et al. (2012) reported that they based their sample size on several outcome variables, but did not clearly specify the method that was used.

3.2.2 Trials with co-primary outcomes

Seven (3%) of the 209 trials reported co-primary outcomes. Even though it is unnecessary, two of these trials accounted for multiple comparisons of the co-primary outcomes in their analysis: one used the Hochberg method and one used a pre-specified testing hierarchy. The abstracts of the seven trials clearly specified all the co-primary outcomes. When calculating the sample size, five of the trials performed calculations based on all co-primary outcomes and two of the trials performed calculations based on just one outcome.

3.2.3 Trials with one stated primary outcome

The remaining 142 (67%) trials reported that only one of the multiple outcomes was primary. Of these trials: five reported one primary and no secondary outcomes; and another six (3%) used a composite primary outcome.

3.2.4 Psychiatry, neurology and general medicine journals

The results were also reviewed by the disease area of the journal in which the article was published. The journals were grouped into the areas: psychiatry, neurology and general medicine. Of these disease areas, the psychiatric journals reported multiple primary outcomes the most frequently with 35 (47%) of the trials reporting multiple primary outcomes. This compared to 18 trials (22%) in the neurological journals and 7 (13%) in the general medicine journals reporting multiple primary outcomes.

Of those trials analysing multiple primary outcomes, 27/35 (77%) and 15/18 (83%) of the trials in the psychiatric and the neurological journals respectively did not account for multiplicity compared to 3/7 (43%) trials in the general medicine journals.

3.2.5 Drug versus non-drug trials

134 (64%) trials evaluated drug treatments, of which 30 (22%) reported multiple primary outcomes of which only six (20%) accounted for multiplicity. Whereas 30 (40%) of the 75 non-drug trials analysed multiple primary outcomes of which nine (29%) accounted for multiplicity.

3.2.6 Secondary outcomes

Nineteen (13%) of the 142 trials that reported only one outcome as primary accounted for multiplicity in their secondary outcomes by adjusting the p-values, even though adjustments may be less important for secondary outcomes (Moyé, 2003). An additional seven (5%) trials

highlighted a main secondary outcome. Five of the 67 (7%) trials that had multiple primary outcomes or co-primary outcomes adjusted for multiple comparisons in the analysis of their secondary outcomes: two trials used sequential testing and three trials adjusted the p-values.

Some of the investigators that used only one primary outcome highlight their awareness of the problems associated with multiplicity by stating they did not adjust for multiple *secondary* outcomes. Weaver et al. (2012) said that "no formal correction for multiple analyses" were made. Nobile-Orazio et al. (2012) stated that "despite the large number of tests done on the secondary outcomes, the type I error rate was not adjusted because these analyses were mainly supportive".

3.3 Case Study

The following section focuses on Hong et al. (2011), a paper included in the review. The investigators did not account for multiplicity in the analysis. The aim of this case study is to view if the conclusions in the article would have changed had multiple outcomes been accounted for in the analysis.

The investigators considered the effects of moderate-dose treatment with varenicline on neurobiological and cognitive biomarkers in smokers and non-smokers with schizophrenia or schizoaffective disorder. The objective of the study was to investigate the effect of varenicline on key biomarkers that are associated with schizophrenia. They stated seven key biomarkers as their primary endpoints: prepulse inhibition, sensory gating, antisaccade, visual spatial working memory, eyetracking, processing speed, and sustained attention. No measures were taken to account for the use of multiple primary outcomes. As stated in the title of the report the investigators considered the effect of the intervention in smokers and non-smokers. This resulted in a large number of results being presented. The investigators appear to selectively report the results with different outcomes being presented differently. The results were either presented individually (smokers and non-smokers) or presented combined (all participants).

Table 3.2 The seven primary end points and corresponding p-values taken from the Hong et al. (2011) manuscript.

The original p-values were not adjusted for multiplicity. The Bonferroni, Holm, Hochberg and Hommel adjustment methods for multiplicity have been applied.

	Outcome	Adjustment method				
		None	Bonferroni	Holm	Hochberg	Hommel
1.	Prepulse inhibition	Not Sig*				
2.	Sensory gating	0.006	0.042	0.042	0.042	0.042
3.	Antisaccade	0.034	0.238	0.204	0.204	0.204
4.	Visual spatial working memory	Not Sig*				
5.	Eyetracking	Not Sig*				
6.	Processing speed	Not Sig*				
7.	Sustained attention	Not Sig*				

*These outcomes were reported as not statistically significant, $p > 0.05$. They will remain non-statistically significant when any is adjustment is applied.

As highlighted in the review, a variety of adjustment methods are used in published randomised trials, so several methods have been used to account for multiplicity in this case study. For any of the selected adjustment methods, the sensory gate outcome remains statistically significant when comparing p=0.042 against the nominal 0.05 significance level. However, the evidence of an effect of the intervention for this outcome has been reduced from strong evidence to moderate evidence. The p-value corresponding to the antisaccade outcome substantially increased to approximately 0.2 (ranging from 0.204 to 0.238). If any of the adjustment methods had been used, the conclusion drawn would have been that there is no evidence of an effect of the intervention for this outcome. All other outcomes would remain not statistically significant. This case study demonstrates the importance of adjusting the p-values, or equivalently the significance level, to account for multiple primary outcomes.

3.4 Discussion

The review performed in this chapter has identified that multiple primary outcomes are commonly reported and analysed in RCTs that were published in high impact research journals. It was found that there is a lack of consistency in the reporting and analysis of the outcomes. It was often difficult to determine from the report the number of primary

outcomes being analysed in a trial and there was a lack of consistency when specifying the primary outcome.

A variety of methods to handle multiple primary outcomes were noted in this review. The majority of authors who accounted for multiplicity did so by adjusting the p-values. The most common technique observed was the Bonferroni method. If the outcomes are correlated, this adjustment method is conservative.

In recent years, more complex multivariate methods have been used that utilise the positive-pairwise correlations between outcomes (Teixeira-Pinto et al., 2009). The MANOVA method was the only multivariate method used despite the fact that multivariate methods could increase the power (Teixeira-Pinto et al., 2009, Yoon et al., 2011).

In the majority of cases, the trials did not specify any steps to safeguard the inferences made when using multiple primary outcomes. Of these trials, 26 reported significant results. However, six of these would have drawn different conclusions if the Bonferroni method had been applied. In one trial, the intervention would not have been reported as effective for any of the primary outcomes and in five trials the intervention would be reported as effective for a small subset of the primary outcomes.

Paradoxically, multiple authors demonstrated their awareness of the problems associated with multiplicity by stating they did *not* adjust for multiple outcomes. One reason given for not using any adjustments was “to prevent type II error” (Vitiello et al., 2014) whilst others did not provide any justification.

The proportion of studies not adjusting for multiple outcomes may be underestimated in this review due to selective outcome reporting. A review of trials highlighted that selective reporting of outcomes, where only a subset of original outcome measures are fully reported, frequently occurs in randomised controlled trial (Dwan et al., 2013, Sendyk et al., 2019) . In some instances, multiple outcomes are pre-specified but only a subset are reported as primary outcomes. Consequently, the proportion of trials that fail to address the issues associated with multiplicity may be greater than observed here due to biases in reporting.

For many of the trials, the authors correctly identified the primary outcomes in the abstract, however, there is still considerable room for improvement. For instance, a number of abstracts provided incomplete descriptions. Some abstracts also discussed multiple outcomes without any distinction between the primary and secondary outcomes.

The review performed here focuses on neurology and psychiatry trials although the review is likely to have wider applicability as multiple primary outcomes are also common in other disease areas. The review was also restricted to high quality, high impact journals. It is expected that the proportion of trials that fail to address the issues associated with multiplicity is actually greater in lower impact journals.

3.5 Conclusions

For the neurology and psychiatry RCTs considered in this chapter, which were published in a number of leading medical journals, it was found that multiple primary outcomes were commonly used but often inadequately analysed. More complex multivariate methods could have been used that utilise the pairwise correlations between outcomes. A comparison of the complex multivariate methods would be beneficial to allow recommendations of methods to use in future trials.

Chapter 4

Methods to adjust for multiple primary outcomes in the analysis and sample size calculation of randomised controlled trials

It has been shown that multiple primary outcomes are commonly analysed to characterise the effect of an intervention in RCTs and it is common for these outcomes to be correlated. To investigate whether the intervention is effective for each outcome, it is necessary to perform multiple statistical tests. As discussed in previous chapters, when performing these tests, it is important to control the family wise error rate (FWER) at the nominal significance level. A common approach is to adjust the p-values produced by each statistical test. A variety of methods to adjust the p-values were reviewed in Chapter 2. In clinical trials, it is also important to consider the power of the tests to detect an effect of the intervention. When there are multiple outcomes, the power of the study can be defined in a number of ways depending on the clinical objective of the trial. First, the disjunctive power is the probability of finding at least one true intervention effect across all of the outcomes (Bretz et al., 2010). Second, the conjunctive power is the probability of finding a true intervention effect for all outcomes. Lastly, the marginal power is the probability of finding a true intervention effect for a particular outcome and is calculated separately for each outcome. To investigate multiple primary outcomes, we are typically interested in the disjunctive and marginal power, as recommended by Dmitrienko et al. (2009).

The power requirements of a trial should match the clinical objectives, which need to be specified when designing the study. The sample size calculation should be calculated according to the clinical objectives. It was shown in my review of published RCTs (Chapter 3) that in current practice the sample size calculations in trials often focuses on the marginal power for each outcome. An approach that has been recommended and is often used in trials is to calculate the sample size separately for each of the primary outcomes by applying the Bonferroni method to amend the significance level (Chow et al., 2017). The largest value of the sample size is then considered as the final sample size for the trial (Odekerken et al., 2012).

As previously mentioned, missing outcome data are common in RCTs which will inevitably reduce the power and efficiency of the study (Bell et al., 2014). As a result, there may be failure to detect true intervention effects, when they are present. As such, when considering the methods that adjust for multiple primary outcomes it is also important to consider the consequences and impact of missing outcome data.

When using multiple primary outcomes, there is limited guidance as to which method(s) should be used to take account of multiplicity especially when there are missing data in the primary outcomes. Guidance is needed for both the sample size calculation and the statistical analysis of RCTs with multiple outcomes.

Some studies have compared a selection of methods that adjust p-values to account for multiplicity to handle multiple outcomes in trials. Sankoh, Huque and Dubey (Sankoh et al., 1997) compare a selection of adjustment methods for statistical analysis in terms of FWER but they do not evaluate the methods with respect to the power obtained. Blakesley et al. (2009) discuss both FWER and power requirements for selected methods for a large number of outcomes with varying degrees of correlation. Lafaye de Micheaux et al. (2014) provide formulae to calculate the power and sample size for multiple outcomes. These require several assumptions to be made about the outcomes, including normality and whether the covariance matrix between the outcomes is known or not. They discuss global testing procedures, including the Hotelling T^2 method. None of these studies have investigated the adjustment methods in the presence of missing data.

There is limited literature discussing the sample size requirements for clinical trials with multiple primary outcomes where the clinical objective is to detect an intervention effect for at least one of the outcomes. Dmitrienko et al. (2009) and Senn and Bretz (2007) provide some discussion regarding the sample size in the context of multiple outcomes. However, neither discuss the sample size in the context of which adjustment method should be used. Moreover, they do not provide a comparative table depending on the type of desired power to show implications on the required sample sizes.

In this chapter, I compare methods to adjust p-values in terms of FWER and power. I investigate two and four outcomes when there is complete outcome data and when the outcome data has missing values. I focus on two and four outcomes as my review found that the majority of the trials had considered just two primary outcomes. Additionally, it has been recommended that a trial should have no more than four primary outcomes (Capizzi and Zhang, 1996). I also consider a range of correlations between the outcomes. I consider both

marginal and disjunctive power. Based on my findings, I provide practical recommendations on the adjustment methods which could be used for the sample size calculation and analysis of RCTs with multiple primary outcome. I also present tables showing the implications of using the marginal and disjunctive power on the required sample size for a trial under different scenarios. The work in this chapter has been published in BMC Medical Research Methodology (Vickerstaff et al., 2019). The paper is provided in Appendix 3.

4.1 Aim

The aim of this chapter is to evaluate the validity of selected methods to account for potentially correlated multiple primary outcomes in the analysis and sample size calculation of RCTs.

This aim is split into four objectives. The first objective is to compare methods that account for multiple primary outcomes using the ProCEED case study. The second objective is to perform a simulation study to compare methods to adjust for multiple outcomes in terms of FWER, disjunctive power and marginal power when investigating two or four correlated outcomes. The third objective is to compare the sample size needed to achieve the required marginal and disjunctive power. The final objective is to provide guidance as to which method(s) should be used during the design and analysis of RCTs with multiple primary outcomes which are correlated.

Methods to account for multiple outcomes

The following methods that account for multiplicity are compared in this chapter: Bonferroni, Holm, Hochberg, Hommel, Dubey-Armitage-Parmar (D/AP) and stepdown MinP resampling method.

My review showed that the majority of trials that used multiple outcomes analysed the outcomes separately without any adjustments for multiple comparisons (Chapter 3). When adjustment methods were used, only the most basic methods were used, potentially due to how easily they can be implemented. The Bonferroni method was the most commonly used method, although the Holm and Hochberg methods were also used. As a consequence, in this chapter, I am focusing on the more basic techniques.

The Bonferroni and Holm methods are used as they are well-known methods that are often used in clinical trials. When outcomes are independent, it has been shown that there may be

a gain in power when using the Hochberg and Hommel methods compared to using the Bonferroni and Holm methods. Consequently, I wish to investigate the performance of these methods for different scenarios, especially when the outcomes are correlated and when there are missing data in the outcomes.

I also consider the Dubey/Armitage-Parmar (D/AP) method and stepdown MinP resampling method as these take account of the correlation between outcomes. The D/AP method was selected as there has been little theoretical work to assess the performance of the D/AP method and in which scenarios it should be used, however, it does lend itself to simulation assessment (Sankoh et al., 1997). The resampling methods have previously been recommended for clinical trials with multiple outcomes, but they are not widely used in practice (Reitmeir and Wassmer, 1999). The stepdown MinP has been shown to perform well when compared to other resampling methods (Li and Dye, 2013) and was therefore investigated in this paper.

4.2 Case study

A case study is presented to demonstrate use of the methods in a clinical trial setting. The ProCEED dataset was described in detail in Section 2.5.

Methods

The outcomes (BDI-II, WSAS and EQ-5D) were analysed separately using linear regression, using the univariate framework. Subsequently, methods were applied to the p-values produced by each statistical test. The p-values adjusted for multiplicity were then compared to the nominal significance level 0.05.

Results

For the BDI-II outcome, the standardised intervention effect was estimated to be 0.189 (95% CI 0.031, 0.347). When applying a Bonferroni correction, a p-value of 0.057 was observed which is above the nominal significance level of 0.05. In comparison, when applying the other correction methods a p-value less than the nominal significance level was observed (Holm p= 0.042, Hochberg p= 0.038; Hommel p=0.038 and D/AP p= 0.029).

For the WSAS outcome, the effect of the intervention was statistically significant irrespective of the adjustment made ($p \leq 0.042$). As a consequence, if investigators require a single intervention effect to be detected to deem the intervention effective, then the intervention

is shown to be effective using all of the methods. However, the strength of evidence depends on the method used.

For the EQ-5D outcome, no effect of the intervention is observed ($p \geq 0.097$ for all adjustment methods). The Bonferroni is the most conservative method which is reflected in the corresponding p-value ($p = 0.291$) which is three times as large as the p-value for all other adjustment methods ($p = 0.097$). In other trials the adjustment method chosen could make a difference to whether the intervention effect is deemed statistically significant or not. The ProCEED trial results and all p-values are summarized in Table 4.1.

Table 4.1 Analysis of the ProCEED dataset (top) and adjusting the resulting p-values to account for multiple comparisons (bottom)

Outcome	N	Mean diff.*	SE*	95% CI*	Mean diff. on original scale	P-value
BDI-II	429	0.189	0.081	(0.031, 0.347)	2.762	0.019
WSAS	428	0.195	0.080	(0.038, 0.350)	2.358	0.014
EuroQol	415	-0.146	0.088	(0.318, -0.026)	3.147	0.097
Adjust p-values to account for multiple comparisons						
	Bonferroni	Holm	Hochberg	Hommel	D/AP	No adjustment
BDI-II	0.057	0.042	0.038	0.038	0.029	0.019
WSAS	0.042	0.042	0.038	0.029	0.020	0.014
EuroQol	0.291	0.097	0.097	0.097	0.097	0.097

*These correspond to standardised intervention effects.

BDI-II = Beck Depression Inventory; CI = Confidence interval; D/AP = Dubey/Armitage-Parmar; Mean diff = mean difference; SE = standard error; WSAS = Work and social activities scale.

For this example, I conclude that the intervention is effective for at least one outcome when using any of the adjustment methods. However, the interpretation of the effect of the intervention for each outcome can vary depending on which method is used. In practice, the choice of the adjustment method may also depend on other factors, such as the availability of simultaneous confidence intervals and unbiased estimates (Paux and Dmitrienko, 2018).

4.3 Simulation study

I used the following model to simulate values for two continuous outcomes $\boldsymbol{Y_i} = (Y_{i,1}, Y_{i,2})$:

$$\boldsymbol{Y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta}_1 x_i + \boldsymbol{\epsilon_i}\,. \tag{4.1}$$

Here x_i indicates whether participant i received the intervention or control, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^T$ is the vector of the effects of the intervention for each outcome, $\boldsymbol{\epsilon_i} = \left(\epsilon_{i,1}, \epsilon_{i,2}\right)^T$ are errors which are realisations of a multivariate normal distribution

$$\boldsymbol{\epsilon_i} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),$$

and $\rho \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. The model was also extended to simulate values for four continuous outcomes. The model was extended such that the correlation between any pair of outcomes is the same. I investigated both equal intervention effect sizes and varying effect sizes across outcomes. For the equal intervention effect sizes, I specified an effect size of 0.35 on all outcomes, that is $\boldsymbol{\beta}_1 = (0.35\,, 0.35)^T$ or $\boldsymbol{\beta}_1 = (0.35\,, 0.35, 0.35, 0.35)^T$ for two and four outcomes respectively. This represents a medium effect size, which reflects the anticipated effect size in many RCTs (Rothwell et al., 2018). For the varying intervention effect sizes, I specified $\boldsymbol{\beta}_1 = (0.2\,, 0.4)^T$ or $\boldsymbol{\beta}_1 = (0.1\,, 0.2, 0.3, 0.4)^T$ for two and four outcomes, respectively. I also investigate the effect of skewed data by generating outcome values with equal intervention effect sizes following a gamma distribution with shape and scale parameter of 2. The gamma distribution is often used to model healthcare costs in clinical trials (Thompson and Nixon, 2005, Nixon and Thompson, 2005) and may also be appropriate for skewed clinical outcomes.

I set the sample size to 260 participants, with an equal number of participants assigned to each arm. This provides 80% marginal power to detect a clinically important effect size of 0.35 for each outcome, using an unpaired Student's t-test and an unadjusted significance level of 0.05. I introduced missing data under the assumption that the data were missing completely at random (MCAR). When simulating two outcomes, 15% and 25% of the observations in outcome 1 and 2 were missing respectively, meaning that on average approximately 4% of the observations would be missing for both outcomes. When simulating four outcomes, 15% of the observations are missing in two outcomes and 25% of the observations are missing in the other two outcomes. This proportion of missingness in

outcomes is often observed in RCTs (Killaspy et al., 2015, Osborn et al., 2015, Hassiotis et al., 2014).

I estimated the FWER and disjunctive power by specifying no effect of the intervention ($\beta_{1j} = 0$) and an effect of the intervention ($\beta_{1j} \neq 0$), respectively, and calculating the proportion of times a significant test results was observed for at least one of the outcomes. The marginal power was similarly estimated but I calculated the proportion of times an intervention effect was observed on the nominated outcome. For each scenario, I ran 10,000 simulations. The simulations were run using R version 3.4.2. The stepdown MinP method was implemented using the NPC package (Caughey and Caughey, 2016).

We calculated the sample size based on disjunctive power using the R package "mpe" (Kohl and Kolampally, 2017) and we calculated the sample size based on the marginal power using the R package "samplesize" (Scherer, 2016). The statistical methodology used for the sample size calculation in these packages is described in Appendix 4.

4.3.1 Results

The Bonferroni and Holm methods lead to the same FWER and disjunctive power when analysing multiple primary outcomes. This is because both methods adjust the smallest p-value in the same way. Similarly, the Hochberg and Hommel methods lead to same FWER and disjunctive power when two primary outcomes are analysed and differences between these methods arise when analysing three or more outcomes.

Family wise error rate, FWER

The FWER obtained when evaluating two and four outcomes are displayed in Figure 4.1 and Figure 4.2, respectively. Following on from the explanation above, the Holm and Hommel methods are not displayed in Figure 4.1 and the Holm method is not displayed in Figure 4.2.

When there is correlation between outcomes ($\rho \geq 0.2$), the D/AP method does not control the FWER. All other adjustment methods control the FWER in all scenarios. The stepdown MinP performs well in terms of FWER. Unlike the other methods, it maintains the error rate at 0.05 as the correlation between the outcomes increases. Differences between the Bonferroni, Hochberg and Hommel methods arise when there is moderate correlation between the outcomes ($\rho \geq 0.4$). The Hommel method provides the FWER which is closest to 0.05 whilst being controlled, followed by Hochberg and then Bonferroni. Very similar

results were observed when the outcomes were skewed, consequently these results are presented in the Appendix 5.

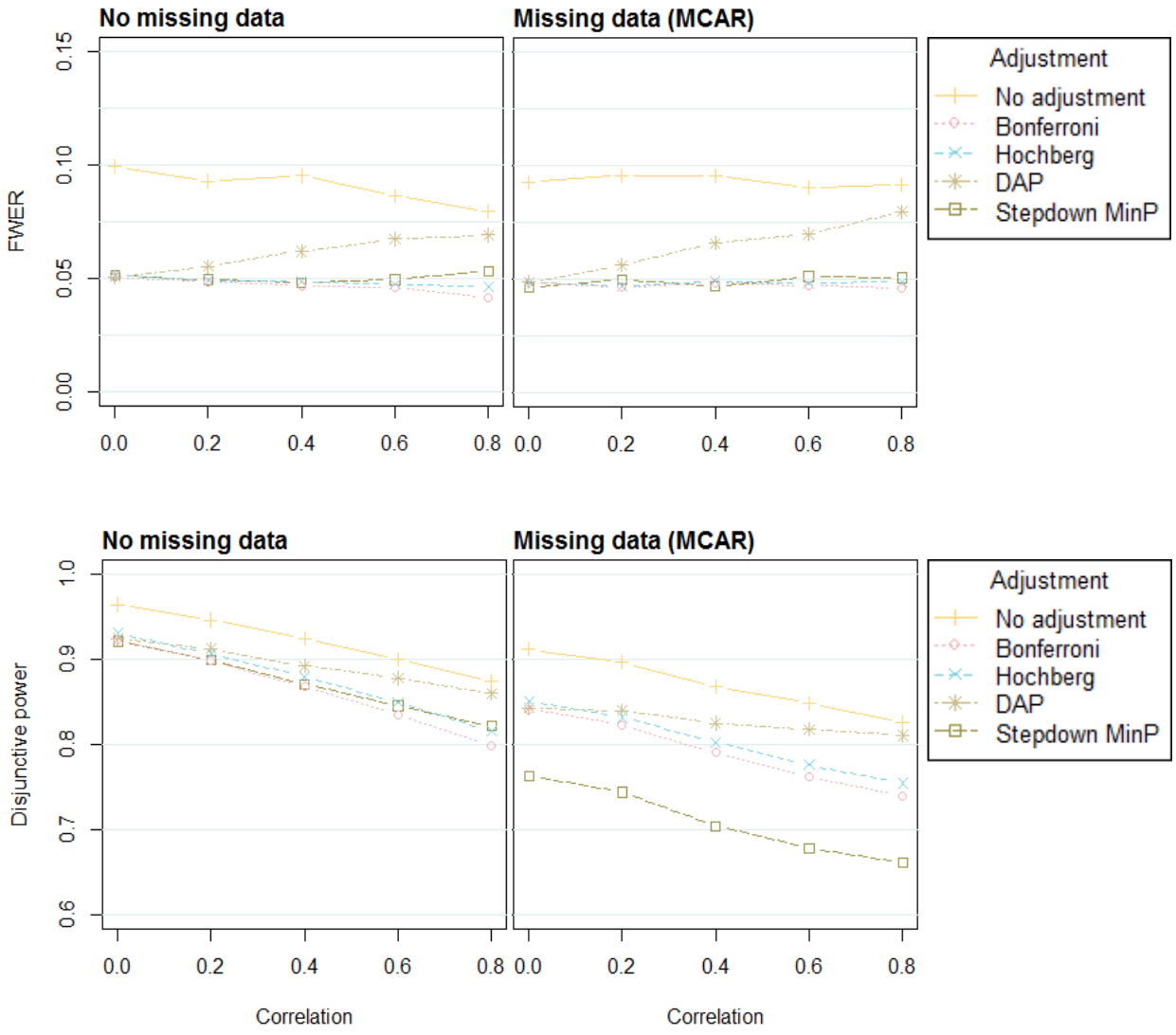
Disjunctive power

Figure 4.1 and Figure 4.2 show that the disjunctive power decreases as the correlation between the outcomes increases for all methods. I did not consider the power obtained when using the D/AP method due to its poor performance in controlling the FWER. When there are no missing data, the stepdown MinP and Hommel methods provide the highest disjunctive power. For weak to moderate correlations ($\rho = 0.2\ to\ 0.6$) the Hommel method has slightly more disjunctive power, but the stepdown MinP method performs better when there is strong correlation ($\rho = 0.8$). The stepdown MinP method gives the lowest power in the presence of missing data. This could be attributed to the fact that it uses listwise deletion removing participants with at least one missing value prior to the analysis which would result in a loss of power due to missing data. As expected the Bonferroni method gives a slightly lower power when compared to the other methods for complete data. However, it considerably outperforms the stepdown MinP method when there is missing data. Very similar results were observed when the outcomes were skewed.

When the intervention effect sizes varied, the differences observed between the methods were less pronounced. When using four outcomes with varying effect sizes, very similar disjunctive power was observed to that of constant effect sizes. These results are displayed in the Appendix 5. When using the Hommel method, higher disjunctive power was observed compared to Holm and Bonferroni methods, albeit by a very minimal amount.

Figure 4.1 The FWER (top) and disjunctive power (bottom) obtained when analysing two continuous outcomes using a variety of methods to control the FWER.

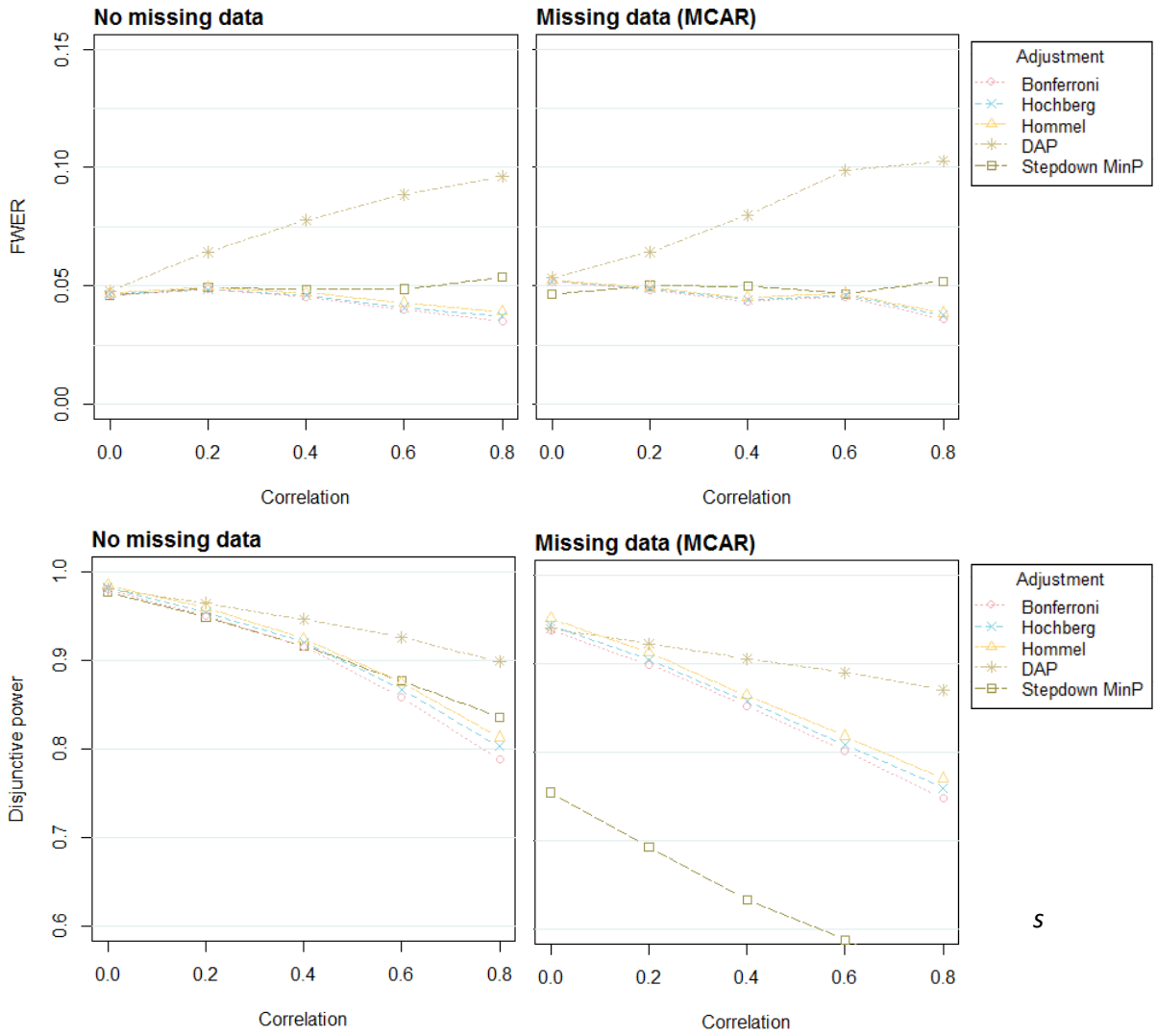
In the left-hand graphs, there are no missing data. In the right-hand graphs, the missing data are missing completely at random, with 15% missing in the first outcome and 25% missing in the second outcome ('Missing data MCAR'). The graphs display various degrees of correlation between the outcomes, range from $\boldsymbol{\rho = 0}$ to $\boldsymbol{\rho = 0.8}$.



*The Monte Carlo standard errors (MCSE) were similar for all methods. When there were no missing data, the MCSE was between 0.002-0.004 for the disjunctive power and 0.002-0.004 for the FWER. In the missing data scenario, the MCSE was between 0.002-0.003 for the disjunctive power and between 0.003-0.005 for the FWER.

Figure 4.2 FWER (top) and disjunctive power (bottom) obtained when analysing four continuous outcomes using a variety of methods to control the FWER.

In the left-hand graphs, there are no missing data. In the right-hand graphs, the missing data are missing completely at random, with 15% missing in two outcomes and 25% missing in the other two outcomes ('Missing data MCAR'). The graphs display various degrees of correlation between the outcomes, range from $\boldsymbol{\rho = 0}$ to $\boldsymbol{\rho = 0.8}$.



*The MCSE were similar for all methods. When there were no missing data, the MCSE was between 0.001-0.004 for the disjunctive power and 0.002-0.004 for the FWER. In the missing data scenario, the MCSE was between 0.001-0.004 for the disjunctive power and between 0.001-0.004 for the FWER.

Marginal power

The marginal power obtained for each outcome when using the different adjustment methods are shown in Table 4.2. In terms of marginal power, the Hommel method was the most powerful method, followed closely by the Hochberg method. When two independent outcomes were analysed, a power of 76.8% was observed after applying a Hommel method.

The power decreased to 75.2% when four outcomes were analysed after applying a Hommel method. As expected the Bonferroni method was the most conservative method, providing the least power. However, contrary to popular belief, the Bonferroni method maintains similar levels of power as the correlation increases.

When analysing two outcomes the proportion of simulations in which an effect of the intervention was observed for neither outcome, one outcome or both outcomes are shown in Table 4.3. When using the Holm method, a statistically significant effect of the intervention was observed on both outcomes for 48%-58% of the simulations. This reduced to 36%-48% of the simulations when using the Bonferroni method. As expected, when using the Hochberg and Hommel method the same results were observed. Compared to the Holm method, simulations with two statistically significant intervention effects were observed more frequently when using the Hochberg and Hommel methods.

Table 4.2 Marginal (individual) power obtained for each outcome, when analysing two outcomes (top) and four outcomes (bottom), using a variety of methods to control the FWER.

Two outcomes						
Pairwise correlation between outcomes	None	Bonferroni	Holm	Hochberg	Hommel	Stepdown MinP
0	80.9	72.4	78.5	79.2	79.2	78.2
0.2	80.6	71.8	77.8	78.6	78.6	77.7
0.4	80.0	71.3	76.6	77.7	77.7	76.7
0.6	80.0	71.0	76.0	77.4	77.4	76.7
0.8	80.3	71.3	75.6	77.4	77.4	77.2
Four outcomes						
Pairwise correlation between outcomes	None	Bonferroni	Holm	Hochberg	Hommel	Stepdown MinP
0	80.5	62.3	73.2	75.0	75.2	72.7
0.2	80.4	62.3	72.6	74.4	74.8	72.2
0.4	80.6	62.4	72.1	74.1	74.4	72.2
0.6	80.3	62.0	70.7	73.1	73.5	72.3
0.8	80.3	61.9	69.7	73.2	73.6	73.5

*D/AP method was not examined due to the poor performance observed when exploring FWER.

Table 4.3 The percentage of simulations in which an intervention effect was observed for neither outcome, one outcome or both outcomes when analysing two outcomes, using a variety of methods to control the FWER.

Method	Pairwise correlation between outcomes	Number of outcomes an intervention effect was observed for		
		0	1	2
Bonferroni	0	16.1	48.4	35.5
	0.2	18.6	43.2	38.2
	0.4	20.6	37.7	41.7
	0.6	23.4	32.7	43.9
	0.8	26.3	26.3	47.5
Holm	0	16.1	35.6	48.3
	0.2	18.6	31.0	50.4
	0.4	20.6	26.4	53.0
	0.6	23.4	22.0	54.6
	0.8	26.3	16.0	57.7
Hochberg	0	15.1	35.6	49.4
	0.2	17.6	31.0	51.5
	0.4	19.3	26.4	54.3
	0.6	22.0	22.0	56.0
	0.8	24.8	16.1	59.1
Hommel	0	15.1	35.6	49.4
	0.2	17.6	31.0	51.5
	0.4	19.3	26.4	54.3
	0.6	22.0	22.0	56.0
	0.8	24.8	16.1	59.1

Sample size calculation

I recommend that the Bonferroni method is used for the sample size calculation when designing trials with multiple correlated outcomes since it can be applied by adjusting the significance level and it maintains the FWER to an acceptable level (up to a correlation of 0.6 between outcomes). As the Hochberg and Hommel methods are data-driven, it is not clear how these more powerful methods could be incorporated into the sample size calculation unless prior data are available, for example, a preliminary study is performed. Determination of the required sample size may be dependent upon simulation based methods rather than an analytic formula, which can be used for the Bonferroni method (Food and Drug Administration, 2017).

In Table 4.4, I present the sample size needed to achieve 90% disjunctive power for trials with two outcomes for varying degrees of correlations between the outcomes for $\rho = \{0.2, 0.4, 0.6, 0.8\}$. For these calculations, I specified that there is an equal allocation of participants between the intervention arms. More details regarding the sample size calculation using the disjunctive power are provided in Senn and Bretz (2007). In order to calculate the sample size a priori information on the degree of correlation between the outcomes is required. For comparison, I also present the sample size required to obtain 90% marginal power for each outcome. For all calculations, I have used the Bonferroni method to account for multiple comparisons. I provide the required sample sizes when analysing four outcomes in Table 4.4. The table provides sample sizes for varying effect sizes. The top line provides an example sample size calculation for four outcomes where there is a small standardised effect size for each of the four outcomes. In this case, the standardised effect is 0.2 for all outcomes. If there is weak pairwise correlation between all four outcomes ($\rho = 0.2$), 325 participants would need to be recruited into each arm to obtain 90% disjunctive power. As the pairwise correlation increases to $\rho = 0.8$ the required sample size increases to 529. The required sample size to obtain 90% marginal for each outcome in this scenario is 716 participants per trial arm. This is the equivalent number of participants required to obtain 90% disjunctive power if the outcomes are perfectly and positively correlated ($\rho = 1.0$). Consequently, the number of participants required to obtain 90% marginal power is greater than the number of participants required to obtain 90% disjunctive power.

In the fourth line of the table, varying intervention effect sizes are expected across the outcomes. For two outcomes, a small intervention effect was expected ($\Delta = 0.2$) whereas a medium intervention effect size is expected for the other two outcomes ($\Delta = 0.5$). For this example, the required sample size is much smaller if 90% disjunctive power is required. Only 75 participants are needed, per arm, if the pairwise correlation between the outcomes is 0.2. The required sample size increases as the strength of the pairwise correlation increases. When there is strong pairwise correlation ($\rho = 0.8$), 98 participants are required per trial arm. In comparison, if the aim is to achieve 90% marginal power for each outcome the sample size would be much higher; 716 participants would be required for each trial arm to achieve 90% power for the two outcomes with a small intervention effect. The required sample size for the other two outcomes with a large effect size would be 116 participants. However, if the investigators would like to achieve 90% marginal power for all outcomes the largest of these values (i.e. 716) would be required. As shown in the example, the required sample size varies considerably depending on if marginal or disjunctive power is used.

Table 4.4 Sample size required to obtain 90% disjunctive power and 90% marginal power for each outcome when analysing two outcomes, after applying the Bonferroni method.

Standardised effect sizes for each of the two outcomes		Sample size required to obtain 90% DISJUNCTIVE power				Sample size required to obtain 90% MARGINAL power for each outcome	
		Correlation between outcomes					
Outcome 1	Outcome 2	0.2	0.4	0.6	0.8	Outcome 1	Outcome 2
0.2	0.2	402	436	475	522	622	622
0.2	0.3	237	251	264	274	622	278
0.2	0.4	145	150	154	156	622	157
0.2	0.5	96	98	99	100	622	101
0.3	0.3	179	194	211	232	278	278
0.3	0.4	126	135	144	152	278	157
0.3	0.5	89	93	97	99	278	101
0.4	0.4	101	109	119	131	157	157
0.4	0.5	78	84	90	96	157	101
0.5	0.5	65	70	76	84	101	101

Note: Sample sizes provided are required per arm.

Table 4.5 Sample size required to obtain 90% disjunctive power and 90% marginal power for each outcome when analysing four outcomes, after applying the Bonferroni method.

Standardised effect sizes for each of the four outcomes				Sample size required to obtain 90% DISJUNCTIVE power				Sample size required to obtain 90% MARGINAL power for each outcome			
				Correlation between outcomes							
Outcome 1	Outcome 2	Outcome 3	Outcome 4	0.2	0.4	0.6	0.8	Outcome1	Outcome2	Outcome3	Outcome4
0.2	0.2	0.2	0.2	325	382	447	529	716	716	716	716
0.2	0.2	0.3	0.3	189	215	242	270	716	716	319	319
0.2	0.2	0.4	0.4	114	127	129	152	716	716	181	181
0.2	0.2	0.5	0.5	75	82	89	98	716	716	116	116
0.3	0.3	0.3	0.3	145	170	199	235	319	319	319	319
0.3	0.3	0.4	0.4	101	117	133	151	319	319	181	181
0.3	0.3	0.5	0.5	71	80	88	98	319	319	116	116
0.4	0.4	0.4	0.4	82	96	112	133	181	181	181	181
0.4	0.4	0.5	0.5	63	73	84	96	181	181	116	116
0.5	0.5	0.5	0.5	52	61	72	85	116	116	116	116

Note: Sample sizes provided are required per arm.

4.3.2 Discussion

When using multiple primary outcomes in RCTs, it is important to control the FWER for confirmatory phase III trials. One method is to adjust the p-values produced by each statistical test for each outcome. Additionally, some of the outcomes are likely to have missing values. Consequently, any potential missing data should be considered when choosing an appropriate method to adjust the p-values.

Statistical Analysis

I found that all of the methods investigated, each controlled the FWER with the exception of the D/AP method. The finding is consistent with the results in Blakesley et al. (2009). The stepdown MinP performed best in terms of FWER. It maintained the error rate at 0.05 as the correlation between the outcomes increases; however, the R package used to implement the method uses listwise deletion, which removed participants with at least one missing value before the analysis resulting in a loss of power. The validity of this method depends on how the method is implemented and the extent of the missing data.

I recommend that the Hommel method is used to control the FWER, provided that the distributional assumptions are met, as it provides slightly more disjunctive power than the Bonferroni and Holm methods. When using the Hommel method, it is assumed that the distribution of the joint test statistic under the alternative hypothesis is known but not fully specified. For example it is known that the joint distribution of the test statistic under the alternative hypothesis is multivariate normality, but the parameters are not specified. This distributional assumption associated with the Hommel method is not restrictive and is encountered in many multiplicity problems arising in clinical trials (Dmitrienko and D'Agostino, 2013). Even when the data followed a skewed distribution, the Hommel method performed well, showing it may be used to analyse a variety of outcomes, including when the normality assumption is violated.

Given the availability of software packages to implement the more powerful methods, there is little reason to use the less powerful methods, such as the Holm method. For example, the Hommel method can easily be implemented in R or SAS. Despite that the Hommel method is not currently available in Stata or SPSS, the p-values can be readily transferred and adjusted in R. However, if the assumptions cannot be met, the simpler Holm method could be used.

When the intervention effect size varied across outcomes, I found that the differences in disjunctive power between the methods were less pronounced. It appeared that the outcome with the largest effect size 'dominated' the disjunctive power. When the sample size is based on disjunctive power, the outcome with the largest effect size would have high marginal power, whereas the outcome with the smallest effect size would have low marginal power – much below the overall desired level of power. It follows that when investigators are looking for an intervention effect for at least one outcome, it is unlikely that they will see an intervention effect for the outcomes with the smaller effect sizes without seeing an intervention effect on the outcomes with the largest effect size. Consequently, for this scenario, it may be advisable to choose the outcome(s) which is expected to have the largest effect size as the primary outcome(s) and treat the other outcomes as secondary outcomes, however, this decision will need to account for the relative clinical importance of the outcomes. Alternatively, when the intervention effect size varies across the outcomes, investigators may wish to consider 'alpha spending' in which the total alpha (usually 0.05) is distributed or 'spent' across the analyses. For example, for a scenario with two primary outcomes, the outcome which is expected to have the largest effect size may be assigned the majority of the alpha with a small portion of the alpha reserved for the alpha with the smallest effect size.

I appreciate that in practice the choice of adjustment method may also depend on other factors, such as the availability of simultaneous confidence intervals and unbiased estimates of the intervention effects. It is standard practice to report 95% confidence intervals alongside point estimates and p-values. When using multiple primary outcomes, it may be necessary to adjust these confidence intervals so that they correspond to the p-values adjusted for multiplicity. The confidence interval may be easily adjusted when using the Bonferroni or Holm methods using the R package AdjustPvalues (Paux and Dmitrienko, 2018). However, it is not straightforward to adjust the confidence interval when using the Hochberg and Hommel methods. Consequently, the reported confidence intervals may not align with the p-values when these methods are used. In this instance, the conclusions should be based on the p-values and not the confidence intervals (European Medical Agency, 2017). If confidence intervals that correspond to the chosen multiplicity adjustment are not available or are difficult to derive, then it is advised to use simple but conservative confidence intervals, such as those based on Bonferroni method (European Medical Agency, 2017).

It is not necessary to control the FWER for all types of trial designs, for example, for trial designs with co-primary outcomes, where all outcomes have to be declared statistically

significant for the intervention to be deemed successful. In this scenario, no adjustment has to be made to control the FWER and the conjunctive power is used. I have not evaluated the conjunctive power as it is not relevant to the scenarios considered in this chapter. The conjunctive power behaves in reverse to the disjunctive power and is substantially reduced compared to the marginal power. As an illustration, if two independent co-primary outcomes are used and there is a marginal power of 80% for each outcome, the conjunctive power of statistical significance for both outcomes is 80% x 80% = 64%. Additionally, as the correlation between outcomes increases, the conjunctive power increases. The conjunctive power will never be larger than the marginal power (80% for this example) and the sample sizes for clinical trials will have to be adjusted accordingly (Senn and Bretz, 2007). The sample size will need to be adjusted to take account of the multiple co-primary outcomes. Formulae and corresponding sample size tables are given in Sugimoto et al. (2012).

Additionally, adjustments for multiple comparisons may not be necessary for early phase drug trials. However, it is generally accepted that adjustments to control the FWER are required in confirmatory studies, that is when the goal of the trial is the definitive proof of a predefined key hypothesis for the final decision making (Bender and Lange, 2001).

My review of trials with multiple outcomes showed that majority of the trials analysed the outcomes separately without any adjustments for multiple comparisons. Where adjustment methods were used, only the most basic methods were used, possibly due to their ease of implementation. The Bonferroni method was the most commonly used method, although the Holm and Hochberg methods were also used. As a consequence, I focused on relatively simple techniques in this chapter. However, more advanced methods, such as graphical methods to control the FWER are available and described in Bretz et al. (2011) and Bretz et al. (2009).

Regardless of the adjustment method chosen, for all trials using multiple outcomes, the analysis plan should clearly describe how the outcomes will be tested including which adjustment method, if any, will be used (Food and Drug Administration, 2017). Any additional outcomes that have not been pre-specified in the analysis plan cannot, in general be used to demonstrate the effectiveness of an intervention, even in successful trials (Food and Drug Administration, 2017).

Sample size

When designing a clinical trial, it is important to calculate the sample size needed to detect a clinically important intervention effect. Usually the number of participants that can be recruited in a trial is restricted because of ethical, economic and practical considerations. However, if the sample size is too small it may not be possible to detect an important effect. The sample size calculation is usually based on an appropriate statistical method which will be used for the primary analysis depending on the study design and objectives. The required sample size can vary greatly depending on if marginal power or overall disjunctive power is used, which highlights the importance of considering the clinical objective of the trial in the sample size calculation. To account for multiplicity in the sample size calculation, I recommend that the Bonferroni method is used. The Bonferroni method can be applied easily within the sample size calculation using an analytical formula (Food and Drug Administration, 2017) and our simulation study showed that it maintains the FWER to an acceptable level for low to moderate correlation between the outcomes. Additionally, there is only a small reduction in power when using the Bonferroni method when compared to the other methods in the presence of missing data. In contrast, the other methods investigated in this paper are data driven. For these methods, it is unclear how these can be implemented in the absence of a preliminary study.

One method that has been used to calculate the sample size, for multiple primary outcomes, is to calculate the sample size based on the individual marginal power for each outcome and to then choose the maximum sample size for the trial (Odekerken et al., 2012). This method guarantees adequate marginal power for each individual test. However, this approach will overestimate the number of participants required if the investigators are interested in disjunctive power. Moreover, it may be difficult to achieve the required sample size in trials where recruitment is a problem. As such, trials may be closed down prematurely. Finally, I recommend that the sample size should be inflated to account for the expected amount of missing data.

Study extensions and limitations

In this chapter, I only investigated continuous outcomes. However, in randomised controlled trials binary outcomes or a combination of continuous and binary outcomes may be used. For two binary outcomes, the maximum possible pairwise correlation between the outcomes will be less than one in absolute magnitude (Warner, 2008). Therefore, I would expect similar

results for two binary outcomes but with less pronounced differences between methods for the strong correlations.

Additionally, I only investigated global effects, that is either no interventions effect on any of the outcomes ($\beta_{1j} = 0$) or an intervention effect on all the outcomes ($\beta_{1j} \neq 0$). Global effects are most realistic when the strength of the correlation between the outcomes is moderate to strong. However, in practice a mixture of no effects and some intervention effects may be observed, especially when the strength of the correlation between the outcomes is weak.

4.3.3 Conclusions

To ensure that the FWER is controlled when analysing multiple primary outcomes in confirmatory randomised controlled trials, I recommend that either the Hochberg or Hommel method is used in the analysis for optimal power, when the distributional assumptions are met. When designing the trial, the sample size should be calculated according to the clinical objective of the trial. When specifying multiple primary outcomes, if considered appropriate, disjunctive power could be used, which has smaller sample size requirements compared to that when using the individual marginal powers. The Bonferroni method can be used in the sample size calculation to account for multiplicity.

Chapter 5

Evaluation of multivariate methods to analyse multiple outcomes in clinical trials

Several approaches have been used to analyse trials with multiple outcomes in the presence of missing data. A simple approach to the analysis of multiple primary outcomes is to analyse each outcome separately. As found in the review in Chapter 3, this has been the most common approach to analyse multiple primary outcomes in recently published randomised trials. Patients are typically omitted from any analysis for which they have missing outcome data. However, this approach does not account for the correlation between the outcomes and consequently the precision of the estimates and the power may be lower than that achieved by other approaches (Teixeira-Pinto et al., 2009).

In Chapter 2, I reviewed the methods that were recently used in the literature to analyse multiple outcomes. It was noted that multivariate methods make use of the correlations between outcomes and can provide more efficient estimators when some outcomes have missing values. The multivariate models discussed were the factorisation model, the latent variable model and the multivariate multilevel (MM) model. All three models can handle continuous outcomes, binary outcomes or a combination of the two. In addition, these models can handle non-overlapping missingness and therefore the number of observations does not need to be equal across outcomes. The factorisation, latent and MM models can easily be extended to handle several outcomes, although the factorisation model can be cumbersome when there are more than three outcomes. For this reason, I focus on the latent and MM models and set out to investigate the scenarios in which multivariate methods are superior, and to what extent, with respect to the efficiency gained.

For a comparison, I also investigated analysing outcomes separately with and without imputation of missing data values. Complete case analysis is often used in practice, although imputation is recommended to handle missing data prior to analysis. Multiple imputation is a common imputation method that has become readily available in recent years with packages available in most statistical programs, including R, Stata and SAS.

5.1 Aim

The aim of this chapter is to compare the multivariate multilevel (MM) and latent variable (LV) models to univariate models with (MI+UV) and without multiple imputation (UV) with respect to power and FWER. In the trial setting, it is important to have sufficient power to detect the true intervention effects, when they are present, whilst controlling the FWER. Consequently, I focus on the disjunctive power and FWER obtained when using these methods. Recommendations are made regarding which of these methods provides the most power whilst controlling the FWER.

5.2 Methods

Several scenarios were considered by varying the number of outcomes, the outcome type, the correlation between the outcomes, the size of the intervention effect, the missing data mechanism and the percentage of missing data values. Details of the different simulation factors considered are described in Table 5.1.

The following model was used to simulate values for two continuous outcomes $\boldsymbol{Y_i} = \left(Y_{i,1}, Y_{i,2}\right)^T$,

$$\boldsymbol{Y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta}_1 x_i + \boldsymbol{\epsilon_i}\,, \tag{5.1}$$

where x_i indicates whether the participant i received intervention ($x_i = 1$) or control ($x_i = 0$), $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^T$ is the vector of the effect of the intervention for each outcome, $\boldsymbol{\epsilon_i} = \left(\epsilon_{i,1}, \epsilon_{i,2}\right)^T$ are errors which are realisations of a multivariate normal distribution

$$\boldsymbol{\epsilon_i} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),$$

and ρ is the correlation between outcomes. The model was also extended in the obvious way to simulate four continuous outcomes. To simulate binary outcomes a similar model was used,with an extra final step todichotomise the continuous outcomes at zero.

Table 5.1 Scenarios simulated to evaluate methods which may be used to analyse multiple outcomes.

Variable	**Simulation factors**
Number of outcomes	2 or 4
Outcome type	Continuous; binary; and mixed (half continuous and half binary)
Correlation between outcomes	0.0, 0.2, 0.4, 0.6, 0.8
Effect size (ES) of intervention effect	Continuous outcomes Equal: ES = 0.35 for all outcomes. Varying ES = $(0.2, 0.4)^T$ or ES = $(0.1, 0.2, 0.3, 0.4)^T$ for two and four outcomes respectively. Binary outcomes Equal: the percentage of events in the control and intervention arms were 50% and 65%, respectively for all outcomes (equivalent to an effect size of 0.385). Mixed outcomes Equal: ES = 0.35 for all outcomes.
Missing data mechanism	Missing completely at random (MCAR), missing at random (MAR)
Percentage of missing data values	Low and high levels of missingness. Percentages varied on depending on the missingness mechanisms and the number of outcomes, as described below: *MCAR and MAR, 2 outcomes* Low: 15% and 25% missing values in outcome 1 and 2 High: 30% and 50% missing values in outcome 1 and 2 *MCAR and MAR, 4 outcomes* Low: 15%, 15%, 25% and 25% missing values in outcome 1, 2, 3 and 4 High: 20%, 30%, 40% and 50% missing values in outcome 1, 2, 3 and 4

The sample size was set as 260 for the continuous and mixed scenarios and 340 for the binary scenarios, with equal numbers of participants being allocated to the two intervention groups. These numbers were obtained from sample size calculations for a single outcome using the equal effect sizes in Table 5.1, 5% statistical significance and 80% power.

Missing data was introduced under a variety of assumptions. I specified two forms of missingness: missing completely at random (MCAR) and missing at random (MAR). For both

MCAR and MAR, I investigate low and high levels of missingness as described in the Table 5.1 above.

I expect that the low levels of missingness are representative of many clinical trials. The high levels of missingness are less frequently observed in real settings and represent more extreme scenarios. However, these extreme levels of missingness do occur in clinical trials, for example, in the 10TT trial described in Chapter 2, only 58% of the randomised participants completed the weight loss outcome at two years and 31% of randomised participants had complete quality of life and cost data at two years (Beeken et al., 2017).

Missingness was implemented by simulating values from a multivariate Bernoulli distribution (Leisch et al., 1998) and setting the outcome variables to be missing depending on the corresponding binary indicator. When data are MCAR the missingness does not depend on the observed outcome value or the intervention arm. In the MAR scenarios the probability of misssingness depends on the intervention group with outcome being more likely to be missing in the control arm compared to the intervention arm.

I estimated the FWER by specifying that the intervention had no effect ($\boldsymbol{\beta_1 = 0}$) then calculating the proportion of times a significant test result was observed for at least one of the outcomes over 10,000 simulations. To control the FWER, the Holm method was used. To estimated the disjunctive power a similar approach was used with a specified intervention effect ($\boldsymbol{\beta_1 \neq 0}$). The bias associated with the estimated intervention effects was calculated as the difference between the average intervention effect $\hat{\beta}$ and the true value of the intervention effect β, as given by

$$Bias = \hat{\beta} - \beta.$$

More specifically, $\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i$ is the sample mean of the estimate of the intervention effect, where N is the number of simulations performed, $\hat{\beta}_i$ is estimated intervention effect within each of the N simulations. The amount of bias that is considered problematic to be has varied from $\frac{1}{2}SE(\hat{\beta})$ (Schafer and Graham, 2002) to $2SE(\hat{\beta})$ (Sinharay et al., 2001), where $SE(\hat{\beta})$ is the empirical standard error (EmpSE) of the estimated effect. The empirical SE estimates the long-run standard deviation of $\hat{\beta}$ over the N simulation repetitions. I consider any bias greater than $\frac{1}{2}SE(\hat{\beta})$ to be as problematic. The Monte Carlo standard error (MCSE) was also calculated to provide an estimate of the inaccuracy of the Monte Carlo samples for each scenario.

The following methods of analysis were used:

1) Univariate models (UV). This was used as the comparator for the other methods.
2) Multiple imputation followed by univariate models (MI+UV).
3) Multivariate multilevel model (MM).
4) Latent variable model (LV).

For the univariate approach, the continuous outcomes were analysed using a linear regression model and the binary outcomes were analysed using a probit regression model. The latter was used as this corresponds to how the data were generated.

Multiple imputation was implemented using chained equations (MICE) since this is one of the most widely used methods to impute missing data (Sterne et al., 2009). The outcomes in the two intervention arms were imputed separately which is equivalent to imputing the missing values conditional on the intervention arm. Forty imputations were used for all scenarios, which is the recommended number of imputations when 50% of the data are missing (Graham et al., 2007). Estimates were pooled across imputed datasets using Rubin's rules (Rubin, 2004). The LV models used adaptive quadrature (Rabe-Hesketh et al., 2005) with ten integration points to fit the models by maximum likelihood estimation. To ensure that the model is not over parameterised so that all model parameters are identifiable, one of the parameters must be fixed. The parameter to be fixed should be specified carefully on a case-by-case basis. For all scenarios in this chapter, the latent factor variance was fixed to 0.8 (Grilli and Rampichini, 2006). In the scenarios with binary outcomes, we fixed the latent factor to 1. The MM model was implemented in MLwiN via R using the package 'R2MLwiN' (Leckie and Charlton, 2013, Zhang et al., 2016b). The MI+UV model was implemented using the 'mice' package in R; and the LV method was implemented using GLLAMM in Stata Release 14 (StataCorp, 2015).

5.3 Results

To begin, I compare each method to the UV model and later I compare all the methods together.

Multiple imputation and then univariate models, MI + UV

The FWER varies between 3.1% and 5.9%. The FWER and disjunctive power obtained when using two continuous outcomes are shown in Table 5.2a. When using continuous outcomes, the estimates of the intervention effects were unbiased (results shown in Appendix 6). When there is weak correlation between the two outcomes ($\rho < 0.4$), the imputed values are highly variable which leads to slightly higher empirical standard errors for the estimated intervention effects compared to using the complete data only. The empirical standard errors for the estimated intervention effects are shown in Appendix 6. As a consequence, the disjunctive power observed when using the MI+UV is *reduced* by up to 17% compared to UV method (results shown in Table 5.2a). In contrast, when there is strong correlation between the outcomes and the missing data are not overlapping across outcomes, if MI+UV approach is used, the observed outcome values are highly predictive of the missing outcome values. This leads to smaller empirical standard errors, as shown in Appendix 6. Consequently, the disjunctive power for MI+UV is *increased* compared to UV, by up to 6%.

When analysing four continuous outcomes, the MI+UV approach performs slightly better. When there is no correlation between the outcomes, the disjunctive power is still reduced compared to the UV, however, by only 2%. Whereas for strong correlation between the outcomes, the disjunctive power for MI+UV is increased compared to UV by 11%.

When analysing two binary outcomes (Table 5.2b), the MI+UV approach had lower disjunctive power when compared to analysing the outcomes separately when there is a low proportion of missing data under both MCAR and MAR scenarios. When analysing two binary outcomes, the FWER is maintained at an acceptable level (FWER $\leq$ 5.1%). A small bias in the estimates of the intervention effects was observed when analysing two binary outcomes (results shown in Appendix 6). This may be due to the multiple imputation program requiring us to use logistic regression for the imputation instead of a probit regression as per the rest of the analyses.

When analysing mixed outcome types, the MI+UV only slightly increases the disjunctive power to detect an intervention effect compared to UV. When there is a large proportion of missing data and there is strong correlation between the two outcomes a 4% disjunctive power gain can be achieved compared to UV.

Table 5.2a FWER and disjunctive power when evaluating two continuous outcomes.
Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of Missingness ↓	% of missing values for each outcome ↓	ρ ↓	Family wise error rate (FWER)				Disjunctive power				Relative power (vs. UV)		
		Method →	UV	MI + UV	MM	LV	UV	MI + UV	MM	LV	MI+UV	MM	LV
Complete	(0%, 0%)	0	0.051	-	0.054	0.053	0.923	-	0.927	0.922	-	1.00	1.00
		0.2	0.048	-	0.050	0.050	0.898	-	0.903	0.903	-	1.01	1.00
		0.4	0.047	-	0.049	0.049	0.868	-	0.872	0.872	-	1.01	1.01
		0.6	0.046	-	0.048	0.048	0.834	-	0.840	0.840	-	1.01	1.01
		0.8	0.041	-	0.044	0.053	0.798	-	0.804	0.809	-	1.01	1.01
MCAR	(15%, 25%)	0	0.049	0.047	0.052	0.051	0.841	0.806	0.849	0.845	0.96	1.01	1.01
		0.2	0.046	0.044	0.051	0.050	0.823	0.805	0.834	0.833	0.98	1.01	1.01
		0.4	0.048	0.046	0.051	0.051	0.791	0.792	0.803	0.802	1.00	1.02	1.02
		0.6	0.047	0.047	0.049	0.049	0.762	0.783	0.789	0.789	1.03	1.04	1.03
		0.8	0.046	0.049	0.047	0.056	0.739	0.770	0.769	0.776	1.04	1.04	1.05
MCAR	(30%, 50%)	0	0.053	0.033	0.058	0.058	0.710	0.554	0.727	0.724	0.78	1.02	1.02
		0.2	0.050	0.033	0.056	0.055	0.704	0.579	0.720	0.719	0.82	1.02	1.02
		0.4	0.049	0.043	0.054	0.053	0.681	0.652	0.710	0.710	0.96	1.04	1.04
		0.6	0.050	0.053	0.052	0.052	0.651	0.683	0.698	0.698	1.05	1.07	1.07
		0.8	0.049	0.059	0.052	0.062	0.642	0.701	0.698	0.705	1.09	1.09	1.10
MAR	(15%, 25%)	0	0.048	0.043	0.052	0.051	0.839	0.798	0.846	0.843	0.95	1.01	1.01
		0.2	0.047	0.046	0.052	0.052	0.814	0.789	0.825	0.825	0.97	1.01	1.01
		0.4	0.050	0.050	0.053	0.053	0.794	0.798	0.810	0.810	1.01	1.02	1.02
		0.6	0.048	0.051	0.050	0.050	0.766	0.785	0.792	0.792	1.03	1.03	1.03
		0.8	0.043	0.046	0.044	0.052	0.738	0.771	0.769	0.774	1.05	1.04	1.05
MAR	(30%, 50%)	0	0.052	0.031	0.057	0.056	0.709	0.538	0.725	0.723	0.76	1.02	1.02
		0.2	0.051	0.032	0.057	0.056	0.686	0.558	0.707	0.706	0.81	1.03	1.03
		0.4	0.049	0.039	0.052	0.052	0.678	0.636	0.704	0.703	0.94	1.04	1.04
		0.6	0.051	0.053	0.052	0.052	0.658	0.676	0.695	0.695	1.03	1.06	1.06
		0.8	0.048	0.056	0.049	0.058	0.640	0.689	0.689	0.696	1.08	1.08	1.09

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; $\rho *$ = correlation between outcomes. Note: Monte Carlo standard errors (MCSE) were consistent across methods. MCSE Range for FWER = (0.0020, 0.0030); MCSE Range for Power = (0.0027, 0.0049).

Table 5.2b FWER and disjunctive power when analysing *two binary outcomes.*

Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	Family wise error rate (FWER)				Disjunctive power				Relative power (vs. UV)		
		Method →	UV	MI + UV	MM	LV	UV	MI + UV	MM	LV	MI+UV	MM	LV
Complete	(0%, 0%)	0	0.050	-	0.046	0.051	0.914	-	0.913	0.924	-	1.00	1.01
		0.2	0.050	-	0.046	0.051	0.906	-	0.905	0.903	-	1.00	1.00
		0.4	0.055	-	0.049	0.051	0.884	-	0.883	0.886	-	1.00	1.00
		0.6	0.047	-	0.043	0.050	0.868	-	0.867	0.861	-	1.00	0.99
		0.8	0.049	-	0.044	0.038	0.833	-	0.831	0.819	-	1.00	0.98
MCAR	(15%, 25%)	0	0.053	0.051	0.051	0.050	0.843	0.832	0.842	0.848	0.99	1.00	1.01
		0.2	0.048	0.044	0.048	0.046	0.830	0.823	0.829	0.826	0.99	1.00	1.00
		0.4	0.047	0.045	0.046	0.046	0.816	0.811	0.819	0.816	0.99	1.00	1.00
		0.6	0.049	0.044	0.048	0.041	0.793	0.794	0.803	0.789	1.00	1.01	0.99
		0.8	0.045	0.043	0.048	0.036	0.768	0.774	0.786	0.763	1.01	1.02	0.99
MCAR	(30%, 50%)	0	0.048	0.042	0.046	0.044	0.731	0.706	0.730	0.714	0.97	1.00	0.98
		0.2	0.051	0.044	0.051	0.045	0.714	0.692	0.714	0.696	0.97	1.00	0.97
		0.4	0.048	0.045	0.049	0.040	0.706	0.685	0.711	0.680	0.97	1.01	0.96
		0.6	0.049	0.045	0.048	0.035	0.678	0.665	0.693	0.661	0.98	1.02	0.97
		0.8	0.051	0.042	0.048	0.033	0.671	0.666	0.697	0.632	0.99	1.04	0.94
MAR	(15%, 25%)	0	0.049	0.047	0.050	0.050	0.844	0.835	0.844	0.845	0.99	1.00	1.00
		0.2	0.049	0.046	0.048	0.048	0.829	0.822	0.829	0.821	0.99	1.00	0.99
		0.4	0.051	0.048	0.051	0.043	0.812	0.805	0.813	0.801	0.99	1.00	0.99
		0.6	0.051	0.049	0.054	0.045	0.793	0.792	0.801	0.789	1.00	1.01	0.99
		0.8	0.044	0.042	0.046	0.037	0.777	0.781	0.789	0.755	1.00	1.02	0.97
MAR	(30%, 50%)	0	0.046	0.041	0.046	0.046	0.712	0.690	0.711	0.711	0.97	1.00	1.00
		0.2	0.049	0.044	0.048	0.049	0.708	0.687	0.708	0.689	0.97	1.00	0.97
		0.4	0.050	0.045	0.050	0.043	0.693	0.673	0.695	0.676	0.97	1.00	0.98
		0.6	0.050	0.044	0.048	0.039	0.679	0.672	0.689	0.651	0.99	1.01	0.96
		0.8	0.050	0.042	0.051	0.031	0.659	0.664	0.684	0.608	1.01	1.04	0.92

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; ρ = correlation between outcomes. Note: Monte Carlo standard errors (MCSE) for were consistent across methods. MCSE Range for FWER = (0.0016, 0.0023); MCSE Range for Power = (0.0027, 0.0049).

Table 5.2c FWER and disjunctive power when analysing one continuous and one binary outcome ('mixed' outcome type).
Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	Family wise error rate (FWER)				Disjunctive power				relative power (vs. UV)		
		Method →	UV	MI + UV	MM	LV	UV	MI + UV	MM	LV	MI+UV	MM	LV
Complete	(0%, 0%)	0	0.047	-	0.048	0.048	0.855	-	0.858	0.858	-	1.00	1.00
		0.2	0.044	-	0.044	0.044	0.836	-	0.840	0.838	-	1.00	1.00
		0.4	0.043	-	0.045	0.046	0.813	-	0.818	0.815	-	1.01	1.00
		0.6	0.045	-	0.046	0.041	0.791	-	0.795	0.779	-	1.01	0.98
		0.8	0.048	-	0.049	0.030	0.764	-	0.770	0.721	-	1.01	0.94
MCAR	(15%,25%)	0	0.047	0.047	0.049	0.048	0.778	0.779	0.783	0.782	1.00	1.01	1.01
		0.2	0.049	0.050	0.052	0.051	0.756	0.758	0.764	0.763	1.00	1.01	1.01
		0.4	0.047	0.046	0.048	0.046	0.742	0.746	0.754	0.750	1.01	1.02	1.01
		0.6	0.046	0.045	0.049	0.039	0.716	0.723	0.732	0.706	1.01	1.02	0.99
		0.8	0.046	0.043	0.045	0.029	0.693	0.708	0.714	0.671	1.02	1.03	0.97
MCAR	(30%,50%)	0	0.047	0.054	0.049	0.051	0.650	0.660	0.660	0.658	1.02	1.02	1.01
		0.2	0.049	0.052	0.052	0.052	0.641	0.651	0.654	0.651	1.02	1.02	1.01
		0.4	0.047	0.055	0.048	0.049	0.636	0.651	0.652	0.644	1.02	1.03	1.01
		0.6	0.046	0.051	0.049	0.038	0.620	0.643	0.644	0.604	1.04	1.04	0.97
		0.8	0.046	0.047	0.045	0.029	0.609	0.635	0.637	0.595	1.04	1.05	0.98
MAR	(15%,25%)	0	0.051	0.049	0.053	0.053	0.771	0.772	0.778	0.777	1.00	1.01	1.01
		0.2	0.052	0.052	0.053	0.053	0.753	0.752	0.761	0.760	1.00	1.01	1.01
		0.4	0.050	0.050	0.052	0.049	0.731	0.733	0.741	0.736	1.00	1.01	1.01
		0.6	0.052	0.050	0.051	0.042	0.710	0.728	0.732	0.707	1.03	1.03	1.00
		0.8	0.041	0.039	0.042	0.025	0.696	0.712	0.718	0.678	1.02	1.03	0.97
MAR	(30%,50%)	0	0.049	0.052	0.051	0.051	0.645	0.653	0.656	0.655	1.01	1.02	1.02
		0.2	0.051	0.052	0.052	0.051	0.641	0.652	0.654	0.651	1.02	1.02	1.02
		0.4	0.049	0.053	0.053	0.049	0.628	0.648	0.644	0.635	1.03	1.03	1.01
		0.6	0.046	0.050	0.050	0.037	0.614	0.646	0.637	0.603	1.05	1.04	0.98
		0.8	0.051	0.050	0.052	0.032	0.608	0.647	0.639	0.597	1.06	1.05	0.98

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; ρ = correlation between outcomes. Note: Monte Carlo standard errors (MCSE) for the simulation were consistent across methods. MCSE Range for FWER = (0.0019, 0.0028); MCSE Range for Power = (0.0030, 0.0050).

Table 5.3a FWER and disjunctive power when analysing four continuous outcomes.

Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	Family wise error rate (FWER)			Disjunctive power			Relative power (vs. UV)	
		Method →	UV	MI + UV	MM	UV	MI + UV	MM	MI+UV	MM
Complete	(0%, 0%, 0%, 0%)	0	0.046	-	0.051	0.980	-	0.982	-	1.00
		0.2	0.049	-	0.052	0.950	-	0.954	-	1.00
		0.4	0.046	-	0.050	0.915	-	0.920	-	1.01
		0.6	0.040	-	0.043	0.858	-	0.866	-	1.01
		0.8	0.035	-	0.038	0.788	-	0.797	-	1.01
MCAR	(15%, 25%, 15%, 25%)	0	0.052	0.050	0.059	0.937	0.933	0.946	1.00	1.01
		0.2	0.048	0.050	0.056	0.899	0.898	0.907	1.00	1.01
		0.4	0.044	0.049	0.051	0.852	0.864	0.874	1.01	1.03
		0.6	0.045	0.047	0.046	0.801	0.827	0.831	1.03	1.04
		0.8	0.036	0.036	0.036	0.749	0.788	0.789	1.05	1.05
MCAR	(20%, 30%, 40%, 50%)	0	0.045	0.045	0.053	0.876	0.855	0.891	0.98	1.02
		0.2	0.051	0.053	0.057	0.836	0.836	0.859	1.00	1.03
		0.4	0.046	0.053	0.054	0.787	0.811	0.826	1.03	1.05
		0.6	0.047	0.051	0.050	0.739	0.791	0.797	1.07	1.08
		0.8	0.043	0.047	0.040	0.680	0.757	0.750	1.11	1.10
MAR	(15%, 25%, 15%, 25%)	0	0.052	0.050	0.058	0.938	0.931	0.945	0.99	1.01
		0.2	0.048	0.049	0.053	0.902	0.901	0.913	1.00	1.01
		0.4	0.048	0.050	0.053	0.849	0.865	0.874	1.02	1.03
		0.6	0.043	0.047	0.049	0.802	0.829	0.834	1.03	1.04
		0.8	0.039	0.039	0.039	0.748	0.785	0.784	1.05	1.05
MAR	(20%, 30%, 40%, 50%)	0	0.050	0.049	0.059	0.874	0.857	0.891	0.98	1.02
		0.2	0.050	0.053	0.059	0.828	0.824	0.851	0.99	1.03
		0.4	0.048	0.050	0.052	0.783	0.808	0.820	1.03	1.05
		0.6	0.044	0.050	0.049	0.739	0.791	0.798	1.07	1.08
		0.8	0.041	0.043	0.038	0.691	0.763	0.759	1.10	1.10

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; ρ = correlation between outcomes. Note: Monte Carlo standard errors (MCSE) for the simulation were consistent across methods. MCSE Range for FWER = (0.0018, 0.0024); MCSE Range for Power = (0.0013, 0.0047)

Table 5.3b FWER and disjunctive power when analysing *two continuous and two binary outcomes (four 'mixed' outcomes)*.

Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	Family wise error rate (FWER)			Disjunctive power			Relative power (vs. UV)	
		Method →	UV	MI + UV	MM	UV	MI + UV	MM	MI+UV	MM
Complete	(0%, 0%, 0%, 0%)	0	0.048	-	0.050	0.948	-	0.951	-	1.00
		0.2	0.044	-	0.047	0.908	-	0.912	-	1.00
		0.4	0.048	-	0.049	0.874	-	0.878	-	1.01
		0.6	0.040	-	0.041	0.821	-	0.827	-	1.01
		0.8	0.037	-	0.038	0.765	-	0.771	-	1.01
MCAR	(15%, 25%, 15%, 25%)	0	0.050	0.046	0.052	0.883	0.863	0.891	0.98	1.01
		0.2	0.052	0.046	0.056	0.842	0.827	0.852	0.98	1.01
		0.4	0.047	0.044	0.050	0.803	0.801	0.821	1.00	1.02
		0.6	0.044	0.042	0.046	0.755	0.769	0.785	1.02	1.04
		0.8	0.044	0.040	0.045	0.706	0.731	0.749	1.04	1.06
MCAR	(20%, 30%, 40%, 50%)	0	0.050	0.041	0.054	0.811	0.761	0.823	0.94	1.01
		0.2	0.049	0.041	0.052	0.774	0.744	0.796	0.96	1.03
		0.4	0.045	0.038	0.049	0.740	0.730	0.765	0.99	1.03
		0.6	0.046	0.039	0.052	0.703	0.715	0.746	1.02	1.06
		0.8	0.041	0.032	0.042	0.656	0.689	0.712	1.05	1.09
MAR	(15%, 25%, 15%, 25%)	0	0.046	0.042	0.049	0.880	0.856	0.886	0.97	1.01
		0.2	0.051	0.047	0.054	0.841	0.829	0.852	0.99	1.01
		0.4	0.046	0.041	0.046	0.797	0.801	0.819	1.01	1.03
		0.6	0.046	0.044	0.051	0.757	0.773	0.786	1.02	1.04
		0.8	0.040	0.038	0.041	0.711	0.737	0.745	1.04	1.05
MAR	(20%, 30%, 40%, 50%)	0	0.048	0.040	0.054	0.808	0.761	0.820	0.94	1.02
		0.2	0.051	0.043	0.055	0.760	0.728	0.783	0.96	1.03
		0.4	0.048	0.043	0.055	0.738	0.730	0.768	0.99	1.04
		0.6	0.043	0.035	0.045	0.688	0.703	0.731	1.02	1.06
		0.8	0.044	0.036	0.044	0.646	0.688	0.706	1.07	1.09

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; ρ = correlation between outcomes. Note: Monte Carlo standard errors (MCSE) for the simulation were consistent across methods. MCSE Range for FWER = (0.0017, 0.0023); MCSE Range for Power = (0.0020, 0.0048)

Table 5.4a Disjunctive power when analysing two continuous outcomes with varying effect sizes.

Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	Disjunctive Power			Relative power (vs. UV)	
		Method →	UV	MI + UV	MM	MI+UV	MM
Complete	(0%, 0%)	0	0.775	-	0.789	-	1.01
		0.2	0.754	-	0.763	-	1.01
		0.4	0.738	-	0.747	-	1.01
		0.6	0.729	-	0.738	-	1.01
		0.8	0.717	-	0.726	-	1.01
MCAR	(15%,25%)	0	0.641	0.607	0.655	0.95	1.02
		0.2	0.633	0.612	0.650	0.97	1.03
		0.4	0.618	0.629	0.648	1.02	1.05
		0.6	0.601	0.637	0.648	1.06	1.08
		0.8	0.590	0.665	0.666	1.13	1.13
MCAR	(30%,50%)	0	0.475	0.374	0.499	0.79	1.05
		0.2	0.476	0.394	0.508	0.83	1.07
		0.4	0.453	0.435	0.500	0.96	1.10
		0.6	0.442	0.497	0.512	1.12	1.16
		0.8	0.443	0.560	0.551	1.27	1.25
MAR	(15%,25%)	0	0.649	0.612	0.665	0.94	1.02
		0.2	0.630	0.611	0.650	0.97	1.03
		0.4	0.616	0.624	0.644	1.01	1.04
		0.6	0.601	0.638	0.645	1.06	1.07
		0.8	0.592	0.665	0.668	1.12	1.13
MAR	(30%,50%)	0	0.455	0.367	0.480	0.81	1.06
		0.2	0.461	0.383	0.490	0.83	1.06
		0.4	0.444	0.419	0.490	0.95	1.10
		0.6	0.430	0.471	0.496	1.10	1.15
		0.8	0.427	0.553	0.544	1.30	1.28

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; ρ = correlation between outcomes.
Note: Monte Carlo standard errors were consistent across methods. The range of the MCSE was 0.003 to 0.005

Table 5.4b Disjunctive power when analysing *four continuous outcomes with varying effect size.*
Multivariate and univariate methods are compared for scenarios which vary in terms of missingness, percentage of missing data and degree of correlation between outcomes. The Holm method was applied to all scenarios to account for multiplicity.

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	Disjunctive Power			Relative power (vs. UV)	
		Method →	UV	MI + UV	MM	MI+UV	MM
Complete	(0%, 0%, 0%, 0%)	0	0.799	-	0.812	-	1.02
		0.2	0.743	-	0.757	-	1.02
		0.4	0.717	-	0.732	-	1.02
		0.6	0.676	-	0.689	-	1.02
		0.8	0.635	-	0.649	-	1.02
MCAR	(15%, 25%, 15%, 25%)	0	0.652	0.646	0.683	0.99	1.05
		0.2	0.616	0.630	0.646	1.02	1.05
		0.4	0.600	0.620	0.644	1.03	1.07
		0.6	0.558	0.618	0.626	1.11	1.12
		0.8	0.531	0.613	0.619	1.15	1.17
MCAR	(20%, 30%, 40%, 50%)	0	0.510	0.486	0.552	0.95	1.08
		0.2	0.476	0.483	0.532	1.02	1.12
		0.4	0.442	0.498	0.517	1.13	1.17
		0.6	0.423	0.534	0.538	1.26	1.27
		0.8	0.386	0.572	0.553	1.48	1.43
MAR	(15%, 25%, 15%, 25%)	0	0.655	0.648	0.678	0.99	1.03
		0.2	0.619	0.626	0.648	1.01	1.05
		0.4	0.590	0.622	0.636	1.05	1.08
		0.6	0.552	0.613	0.621	1.11	1.13
		0.8	0.517	0.606	0.605	1.17	1.17
MAR	(20%, 30%, 40%, 50%)	0	0.484	0.479	0.528	0.99	1.09
		0.2	0.459	0.479	0.508	1.04	1.11
		0.4	0.437	0.500	0.510	1.15	1.17
		0.6	0.415	0.513	0.519	1.24	1.25
		0.8	0.376	0.555	0.543	1.48	1.44

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; ρ = correlation between outcomes.

Multivariate multilevel model (MM)

The FWER fluctuated around 5%. The highest level of FWER observed when analysing two continuous outcomes was 5.8%. The FWER was highest in the scenarios when there were high levels of missing data.

If there are no missing data, the MM model performed similarly to analysing the outcomes separately. The small difference in FWER and disjunctive power when analysing continuous outcomes may be attributed to the fact the UV p-values are calculated using a Student's t-distribution, whereas the MM p-values are calculated using a normal distribution. The effects of the intervention were unbiased when analysing two continuous outcomes.

As expected, benefits in terms of disjunctive power are seen when there are missing data as the MM model is able to use observations where one of the outcome values is missing. Even when there is weak correlation, the MM model performs better than the UV model. For continuous outcomes, when there is a low proportion of missing data small power gains may be observed when the correlation is strong ($\rho > 0.4$). When the correlation is 0.8, a relative gain of up to 4% may be observed between the MM model and UV models. When there is a large proportion of missing data, up to a 9% gain in disjunctive power was achieved by the MM model compared to the UV model. When analysing four continuous outcomes, similar results are observed. The relative gains between the MM and UV models range from 5%, when there are low levels of missing data, to 10% when there are high levels of missing data and the between outcome correlation is strong. These results are displayed in Table 5.3a. Additionally, similar results are observed when varying intervention effect sizes are used. When analysing two outcomes with varying intervention effect sizes, with high levels of missing data and strong correlation between the outcomes, a gain of up to 10% may be observed between the MM and UV models. These results are displayed in Table 5.4a.

For two binary outcomes, the UV and MM models perform identically when there are no missing data. When there is a low level of missing data (20%) the differences between the MM and UV models are minimal with differences ranging from 0% to a 2% relative increase in disjunctive power. Larger differences are seen when there is a large amount of missing data and the two outcomes are strongly correlated.

For two binary outcomes, the estimates of the intervention effects are unbiased. The largest relative disjunctive power increase, compared to the UV model, is 4% when there is 40% overall missingness and the correlation is strong (ρ=0.8). When analysing two binary outcomes the MM model occasionally did not converge. This most frequently occurred when

the correlation between the two outcomes was strong (ρ= 0.8) and there was no effect of the intervention.

When analysing four binary outcomes, the MM model often did not converge. For example, when simulating no effect of the intervention and no missing data 37.1% of the simulations (n=18558/50000) reported an error and the results were displayed as "NA". Other simulations reported final values but they do not appear to have converged as the results are much larger than expected, for example the coefficients for the estimated effect size are greater than 1000, when I simulated an effect size of 0. Consequently, I have not reported results for four binary outcomes. It is perhaps an unusual scenario that a clinical trial would have four binary outcomes without any continuous outcomes and therefore this should not affect the conclusions regarding the MM model.

Latent variable model (LV)

When using the latent variable model, the FWER observed ranged from 2.5% up to 6.2% across the continuous and mixed scenarios. The results for the model were heavily dependent on the assumptions made regarding the variance of the latent factor.

For two continuous outcomes, the power gains were comparable to that of the MM model. For two mixed outcomes, after fixing the variance of the latent factor, the FWER was overly conservative when the correlation between the outcomes was strong. This resulted in a loss of disjunctive power compared to the other methods.

The use of the LV model was not investigated for four outcomes due to the increased FWER when using two outcomes.

Comparison of methods

When applying the MM, MI+UV and LV model, and using the Holm method to account for multiplicity, the FWER fluctuates around 5% (between 3.1% and 6.2%). In terms of disjunctive power, the MM model performs better than using MI+UV method in the majority of scenarios. When there is little correlation between the two outcomes, the MM model provides a small increase in disjunctive power compared to analysing the outcomes separately. Whereas when using the MI+UV approach, the disjunctive power is decreased as the standard errors are increased. As the correlation increases between two outcomes, the benefits of the MM model continue to increase. When the correlations between the outcomes are very strong ($\rho = 0.8$) the MM model and MI+UV approach perform similarly.

For two continuous outcomes, unbiased estimates of the intervention effect were obtained using all methods. The Monte Carlo standard errors (MCSE) of the disjunctive power and FWER estimates were similar for all methods. For the FWER estimates the MCSE ranged from 0.0020 to 0.0030 and for the disjunctive power estimates the MCSE ranged from 0.0027 to 0.0049. Similar MCSE were found for the analysis of two binary outcomes and mixed outcome types too (as reported in tables 2b and 2c). For binary outcomes, slightly biased estimates of the intervention effect were obtained when using the MM method.

5.4 Case studies

Re-analysis of the ProCEED trial and 10TT trial.

The two real datasets, ProCEED (Buszewicz et al., 2016) and 10TT (Beeken et al., 2012, Beeken et al., 2017), are re-analysed to illustrate the differences and similarities between the multivariate multilevel model (MM) and analysing outcomes separately. The ProCEED dataset includes three continuous outcomes whereas the 10TT dataset includes a combination of continuous and binary outcomes. The code used to implement the MM, using Stata, R and MlwiN are described in Appendix 7.

5.4.1 Pro-active Care and its Evaluation for Enduring Depression Trial, ProCEED

For this analysis, the 24 month outcomes were used. In all analyses, the corresponding baseline values were adjusted for in the model. The outcomes have been standardised so that the estimate of the intervention effect using the three questionnaires can be compared. Standardisation also ensures that no single outcome dominates when using the multivariate technique. On the other hand, standardisation makes the interpretation of data more complex and care is needed when interpreting results of transformed data. For this reason, I also provide results that have been transformed back to the original scales.

When using the MM model, the improvement on the scale is required to be in the same direction for each outcome. That is, there should be a positive correlation between all outcomes. On the WSAS and BDI-II scales, a higher score means greater impairment, whereas on the Euroqol, a lower score means greater impairment. The Euroqol will be reversed to enable the three outcomes to be combined in a multivariate analysis.

The results for the two models are displayed in Table 5.5 (top). The univariate analysis uses complete case analysis, whereas the MM model allow for overlapping missingness. For the

MM model 431 participants are used compared to 429, 428 and 415 participants for the three outcomes when analysing them separately using univariate models. The standard errors are very similar across the models. As multiple tests have been performed, it is important to apply an adjustment to the p-values to control the FWER. The results of applying various adjustments are displayed in Table 5.5 (bottom).

Table 5.5 Analysis of the ProCEED dataset using univariate models and a multivariate multilevel model (top) followed by adjusting the resulting p-values to account for multiple comparisons (bottom)

	N	Mean diff.*	SE*	95% CI*	Mean diff. on original scale	P-value
Univariate analysis						
BDI-II	429	0.189	0.081	(0.031, 0.347)	2.762	0.019
WSAS	428	0.195	0.080	(0.038, 0.350)	2.358	0.014
EuroQol	415	-0.146	0.088	(0.318, -0.026)	3.147	0.097
Multivariate multilevel model analysis						
BDI-II		0.211	0.082	(0.050, 0.372)	3.078	0.010
WSAS	431	0.207	0.081	(0.048, 0.364)	2.500	0.011
EuroQol		-0.146	0.088	(0.318, -0.027)	3.141	0.098
Adjusting the p-values reported above to account for multiple comparisons						
	Bonferroni	Holm	Hochberg	Hommel	D/AP	No adjustment
Univariate analysis						
BDI-II	0.057	0.042	0.038	0.038	0.029	0.019
WSAS	0.042	0.042	0.038	0.029	0.020	0.014
EuroQol	0.291	0.097	0.097	0.097	0.097	0.097
Multivariate multilevel analysis						
BDI-II	0.030	0.030	0.022	0.020	0.014	0.010
WSAS	0.033	0.030	0.022	0.022	0.017	0.011
EuroQol	0.294	0.098	0.098	0.098	0.098	0.098

*Standardised intervention effects.

BDI-II = Beck Depression Inventory; CI = Confidence interval; D/AP = Dubey/Armitage-Parmar; Mean diff = mean difference; SE = standard error; WSAS = Work and social activities scale

In summary, similar results are obtained when using both the MM model and the univariate model. Different conclusions might have been drawn when using the Bonferroni method

compared to using the other adjustment methods, as the p-value adjusted for multiplicity using Bonferroni method increased to above the 0.05 significance level. For all other adjustment methods, the same conclusions should be drawn from both analyses. One advantage of the MM model over the univariate analysis is that a 'composite' joint effect can also be calculated if appropriate. If a joint effect is desired, the investigators would need to decide which of the outcomes to combine. It is possible to combine some of the outcomes into a common effect whilst keeping an individual intervention effect for the remaining outcomes. For example, a joint effect could have been estimated for BDI-II and WSAS whilst simultaneously calculating an individual intervention effect for EuroQol. This would result in less statistical comparisons being performed and, therefore, less stringent rules can be placed on the resulting p-values. In a trial scenario, the decision to combine outcomes would need to be made a priori at the start of the study and documented in the statistical analysis plan.

5.4.2 Ten Top Tips trial

In these analyses, the outcomes were standardised and the corresponding baseline variables were included for in the model. For the univariate analysis, the estimated effects of the intervention on the original scales for the weight and waist circumference effects are -0.872kg and -0.858cm respectively, compared to the MM model results of -0.880kg and -0.888cm, respectively.

After adjusting for multiplicity, the weight outcome remains statistically significant at the 5% level. The most conservative adjustment increases the p-value to 0.012 for both analyses. The waist circumference and glucose level remain above the 5% significance level when any of the adjustments for multiplicity are applied.

The estimated effect for blood glucose differs slightly for the two models. This is likely to be due to missing data for blood glucose, which is ignored by the univariate model. As the MM model uses the correlations between the outcomes, I observed increased disjunctive power and improved precision for the effect of the intervention.

Table 5.6 Analysis of Ten Top Tip dataset using univariate models and a multivariate multilevel model (top) followed by adjusting the resulting p-values to account for multiple comparisons (bottom)

	N	Coef.	Standard error	95% Confidence interval	P-value
Univariate analysis					
Standardised weight	383	-0.052	0.018	(-0.088, -0.016)	0.004
Standardised waist circumference	378	-0.069	.0483	(-0.164, 0.026)	0.153
Blood glucose (normal/high)	330	-0.260	0.314	(-0.875, 0.355)	0.407
Multivariate multilevel analysis					
Standardised weight		-0.053	0.018	(-0.088, -0.017)	0.004
Standardised waist circumference	388	-0.071	0.048	(-0.166, 0.023)	0.138
Blood glucose (normal/high)		-0.295	0.311	(-0.904, 0.315)	0.343

Adjusting the p-values reported above to account for multiple comparisons						
	Bonferroni	Holm	Hochberg	Hommel	D/AP	No adjustment
Univariate analysis						
Weight	0.012	0.012	0.012	0.012	0.007	0.004
Waist circumference	0.459	0.306	0.306	0.306	0.235	0.153
Blood glucose	1.00	0.407	0.407	0.407	0.669	0.407
Multivariate multilevel analysis						
Weight	0.012	0.012	0.012	0.012	0.007	0.004
Waist circumference	0.414	0.276	0.276	0.276	0.213	0.138
Blood glucose	1.000	0.343	0.343	0.343	0.589	0.343

In summary, the MM model allows both continuous and binary outcomes to be analysed simultaneously in a single step. However, I found that in this trial scenario use of the MM model made little difference to the results and conclusions.

5.5 Discussion

In this section, I have performed a simulation study to investigate the differences in disjunctive power and FWER achieved using the multivariate multilevel (MM) model, a latent variable (LV) model and a univariate model with and without multiple imputation (MI+UV and UV, respectively).

The simulations suggested that the power to detect an effect of the intervention can be increased by using multivariate multilevel (MM) models as opposed to analysing each outcome separately with or without multiple imputation (UV and MI+UV). However, I found that the power gains were small in all but extreme scenarios, for example, when there is strong correlation between outcomes or when there are high levels of missing data. Pituch et al. (2016) and Snijders and Bosker (2012) reported efficiency gains for MM model compared to UV models in presence of missing data based on case studies.

When the pairwise correlations between the outcomes were weak, the power was reduced when using the MI+UV approach compared to using the UV approach. These findings are consistent with the results presented in Sullivan et al. (2018), which state that MI may be less efficient than complete case analysis due to Monte Carlo simulation error.

The MM model offers a computational advantage to the MI+UV approach as the MM model enables the analysis to be performed in just one step. In contrast, the MI+UV approach requires three steps: specifying the imputation model and performing the imputation, fitting the analysis model to each imputed dataset; scombining the results across the imputed datasets.

When a single primary outcome is specified in a trial, the MM model can still be used for the analysis of secondary outcomes. Alternatively, when there are missing values in the primary outcome, both the primary and secondary outcomes can be analysed simultaneously using the MM model. Additionally, the MM model allows for joint effects to be estimated although this should be documented in advance in a statistical analysis plan.

The results from the LV model are dependent on the constraints imposed on the model. In this work, the latent factor variance was fixed. For a discussion of alternative constraints see Skrondal and Rabe-Hesketh (2004).

5.6 Conclusions

It was found that the power to detect an effect of an intervention may be increased by using MM models rather than using UV models. However, it was found that the gains were small except in the more extreme scenarios, such as strong correlation between outcomes or high levels of missing data. The MM model may be used as a one-step method instead of the more commonly used MI+UV approach. The MM model may also be useful when analysing multiple correlated secondary outcomes or to estimate a joint intervention effect.

Chapter 6

Evaluation of methods to analyse multiple outcomes when data are missing not at random

The majority of randomised trials have missing outcome data. There is now an understanding that simple approaches, such as discarding the participants with missing data from the analysis is unacceptable (Little et al., 2012). As a consequence, there has been an increase in the use of more complex methods, in particular multiple imputation (MI). As previously discussed in Chapter 2, when implementing these more complex methods, it is usually under the assumption that the underlying missingness mechanism is missing at random (MAR). However, this assumption is untestable, that is, by using the observed data it is not possible to distinguish between MAR and the missingness mechanism missing not at random (MNAR). Misleading inferences and incorrect conclusions may be made if the assumptions about the missingness mechanism are incorrect.

Under the MNAR assumption, parameter estimation from the observed data alone is typically biased. The amount of bias depends on the proportion of dropout and the strength of the relationship between the unobserved outcome and probability of dropout (White and Carlin, 2010). In this chapter, I investigate the whether the multivariate multilevel (MM) model can reduce the bias in the estimated effect of the intervention when the missing data mechanism is MNAR.

6.1 Aim

The aim of this section is to investigate the bias in the estimated coefficients when using the multivariate multilevel model and to compare the results to the bias that arises from analysing the outcomes separately. This includes when multiple imputation is used to handle any missing outcome values that are MNAR.

6.2 Simulation study methods

I generated the data using a similar methodology to that described in the previous chapter. However, for this chapter I investigated scenarios which vary in the number of outcomes,

outcome type, percentage of missing data and degree of correlation between outcomes. The different factors considered are described in Table 6.1.

Table 6.1 Scenarios implemented to investigate methods when missing data are missing not at random.

Factors	
Number of outcomes	2 outcomes, 4 outcomes
Outcome type	Continuous outcomes; binary outcomes; and half continuous and half binary outcomes, which is referred to as 'mixed' outcomes
Pairwise correlation between outcomes	0.0, 0.2, 0.4, 0.6, 0.8
Effect size (ES) of intervention effect	Continuous outcomes Equal: ES = 0.35 for all outcomes. Binary outcomes Equal: the percentage of events in the control and intervention arms were 50% and 65%, respectively for all outcomes. Mixed outcomes Equal: ES = 0.35 for all outcomes.
Missing data mechanism	Missing not at random (MNAR)
Percentage of missing data values	Low: 15% of observations were missing for half the outcomes. The other outcomes had no missing values. Medium: 30% of observations were missing in half the outcomes. The other outcomes had no missing values. High: 50% of observations were missing in half the outcomes, the other outcomes had no missing values. High overlapping: When investigating two outcomes 30% and 50% of observations in outcome 1 and 2 were missing; when investigating four outcomes 20%, 30%, 40%, 50% of observations were missing for each of the outcomes respectively.

Missing data that are missing not at random (MNAR) was introduced with varying quantities of missing data. The percentage of missing data simulated reflects those data observed in published clinical trials (Beeken et al., 2017, Hassiotis et al., 2018, Killaspy et al., 2015). The different scenarios have been referred to as 'low', 'medium', 'high' and 'high overlapping'. The first three scenarios are more realistic when data comes from different sources. For example, the data may be complete when collected from the hospital notes whilst other

patient reported outcomes have a chance of being may be missing. Overlapping missingness is more likely to be observed when all the outcomes are patient reported outcomes.

To simulate the data under the MNAR mechanism, a complete dataset was first simulated. The dataset was then sorted in ascending order according to the outcome in which the missing data was to be introduced. The outcome data were divided into quartile groups and the missingness was introduced in each quartile group. The percentage of values which were set to missing increased for each quartile as shown in Table 6.2.

Table 6.2 The percentage of missing observations per quartile used to simulate data that are missing not at random (MNAR)

	Percentage of observations missing per quartile			
Total percentage missing observations	1st	2nd	3rd	4th
0%	0	0	0	0
15%	0	7.5	22.5	30
20%	0	10	30	40
30%	0	15	45	60
40%	0	20	60	80
50%	0	25	75	100

The following methods of analysis were used:

1) Univariate model (UV). This was used as the comparator for the other methods.
2) Multiple imputation followed by univariate models (MI+UV).
3) Multivariate multilevel model (MM).

The details regarding how these methods were implemented are described in Chapter 5.

The MM, UV+MI and UV methods were compared using the bias, mean square error and coverage of 95% confidence intervals associated with the estimated effect of the intervention (Burton et al., 2006). Assuming, the i^{th} simulated dataset yields a point estimate $\hat{\beta}_i$ with standard error $SE(\hat{\beta}_i)$, for $i = 1, \dots, N$, then the bias is the difference between the average estimate of the intervention effect,

$$\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta}_i,$$

and the true value for the estimate of interest, β, that is

$$Bias = \hat{\beta} - \beta.$$

The mean square error is a measure of accuracy which incorporates both measures of bias and variability and is calculated by

$$MSE = \left(\hat{\beta} - \beta\right)^2 + \left(SE(\hat{\beta})\right)^2,$$

where $SE(\hat{\beta})$ is the empirical standard error of the estimate of the intervention effect over all simulations, $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\beta}_i - \bar{\hat{\beta}})^2}$. The coverage of a confidence interval is the proportion of times that the obtained confidence intervals contains the true specified parameter. In other words, the coverage is the proportion of times the $100(1-\alpha)\%$ confidence interval $\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_i)$ includes β, where $SE(\hat{\beta}_i)$ is the standard error of the estimate of the intervention effect within each simulation and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. The coverage should be approximately equal to the nominal coverage rate, e.g. 95 per cent of samples for 95 per cent confidence intervals, to properly control the type I error rate for testing a null hypothesis of no effect (Collins et al., 2001). If the coverage is too high, it suggests that the confidence intervals are too conservative which could lead to a loss of power. This may be referred to as 'over-coverage'. In contrast, 'under-coverage' is when the coverage is too low. This is unacceptable as it indicates over-confidence which leads to higher than expected type I errors (Burton et al., 2006). One suggested criterion for acceptability of the coverage is that the coverage should not fall outside of approximately two standard errors of the nominal coverage probability (p),

$$(p) = \sqrt{\left(\frac{p(1-p)}{N}\right)}.$$

In this example, I calculate the 95% confidence intervals using 10,000 simulations so the $SE(\hat{p}) = 0.00218$ and hence between 94.56% and 96.44% of the confidence intervals should include the true value.

6.3 Results

Multiple imputation followed by univariate analyses (MI+UV)

The estimated effects of the intervention, and the corresponding biases are shown in Figure 6.1 for varying levels of missing data, when using two continuous outcomes. The corresponding Monte Carlo SE, empirical SE, mean square error and coverage of the 95% confidence intervals associated with the estimated effects of the intervention are provided in Appendix 8.

When analysing two continuous outcomes in which the first outcome is complete and the second outcome has low levels of missing data ('low'), there is a very small amount of bias in the estimates of the intervention effects for both outcomes. This is not problematic given that the bias is within the accepted range of $\pm\frac{1}{2}SE(\hat{\beta})$. The Monte Carlo SE of the estimated bias remained consistent and comparable to those observed when analysing the complete cases, varying between 0.00055 and 0.00062. The empirical standard errors ranged from 0.124 to 0.138. The coverage of the 95% confidence interval for the estimated intervention effect when using the complete case analysis was within the accepted range, varying from 94.6% to 95.1%. This was comparable to using the MI+UV approach in which the coverage varied from 94.5% to 95.0%.

As the amount of missing data increases, higher amount of bias can be observed in the estimates of the intervention effects for outcome 2. When there are high levels of missing data in outcome 2 ('high'), bias may be observed in the estimates of the intervention effects. In this scenario, the empirical SE varied from 0.122 to 0.146. Consequently, any estimated intervention effects below 0.289 can be considered problematic due to high levels of bias. Using the UV method, the estimated intervention effects for outcomes with high levels of missing data are very biased and may be viewed as problematic. When there is no correlation between the two outcomes, the MI+UV approach was unable to reduce the bias in the estimates of the intervention effects compared to only using the UV method. When using the MI+UV approach, the bias in the estimate of the intervention effect decreases as the correlation between outcomes increases. In this scenario, the coverage of the 95% confidence interval for the estimates for outcome 2 was outside of the acceptable range (varying from 63.6% to 71.7% when using the MI+UV approach). This under-coverage is expected given the high levels bias in the estimate of the intervention effect. Outcome 1 in these scenarios does not have any missing data and consequently as expected the estimates of the intervention effect are not biased.

When the outcome variables have overlapping missingness, the MI+UV approach increases the bias in the outcome variable with the least amount of missing data. The values are imputed using the intervention effect estimated from the second outcome which is biased and the intervention group indicator. As the correlation between the outcomes increase the bias in the outcome with the highest level of missing data decreases. The results for varying levels of missing data when using two binary outcomes and one continuous and one binary outcome ('mixed') are shown in Figure 6.2 and Figure 6.3 respectively. A similar situation in terms of bias is observed when analysing binary and 'mixed' outcomes.

The observed bias in the estimated intervention effects for various scenarios when using four continuous outcomes are displayed in Figure 6.4. When analysing four outcomes, the bias is reduced even further. When analysing four outcomes in which two outcomes have 30% missing data ('medium' levels of missing data), no gains in terms of bias may be made when there is no correlation between the outcomes, but when there is moderate correlation between the outcomes ($\rho \geq 0.4$) the bias is reduced compared to using complete case analysis. When there is strong pairwise correlation between the outcomes ($\rho = 0.8$), the MI+UV approach is able to remove the majority of the bias in the estimated intervention effects so that only a small amount of bias is observed.

Multivariate multilevel model, MM

The MM model performs similarly to the MI+UV approach. As displayed in Figure 6.1, when analysing two continuous outcomes and there are low levels of missing data, there is a low level of bias in the estimated intervention effects and the MM model is unable to reduce much of the bias. When both outcomes have missing values ('high overlapping missing'), the gains in terms of bias can be seen when using the MM model compared to analysing the outcomes separately when the correlation between the outcome is at least $\rho = 0.4$. When analysing continuous outcomes, the empirical standard errors were comparable to using UV method. The empirical standard errors of estimate of the intervention effects over all simulations are displayed in Appendix 8.

When analysing two binary outcomes with high levels of overlapping missing data, smaller empirical standard errors were obtained when using the MM model compared to the UV model. However, the standard errors were larger than those seen when using MI+UV approach.

As with the analysis in previous chapters, when analysing four binary outcomes the multivariate model often did not converge. For example, when there is a high level of

overlapping missing data the MM model did not converge in 29% of the simulations. Consequently, I have not reported any results for four binary outcomes. Given it is unusual to have four binary outcomes in a trial, without a continuous outcome, I did not investigate this scenario further.

The results for various scenarios when using two continuous and two binary ('mixed') outcomes are displayed in Figure 6.4, the corresponding empirical standard errors are summarised in Appendix 8. The observed bias in the estimated intervention effects when using the MM model is comparable to that when using MI+UV approach when there is a low to medium amount of missing data. When there are high levels of missing data and weak correlation between the outcome variables ($\rho = 0.2$) small gains in terms of removing the bias in the estimated intervention effects may be seen when using the MM model compared to the UV model. As the correlation between the outcomes increases, larger gains are seen as the bias in the estimated intervention effect decreases when using the MM model.

Figure 6.1 Bias in estimating intervention effect when simulating two continuous outcomes and data are MNAR.

The blue dots represent the average of the estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 1. The red dots represent the average of estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 2. The five dots (of either colour) clustered together represents different correlation between the outcomes from 0 (left) to 0.8 (right in increments of 0.2. Each graph corresponds to adifferent level of missing data as indicated. A Monte Carlo standard error for the estimated bias of between 0.0005 and 0.0007 was observed for all scenarios.

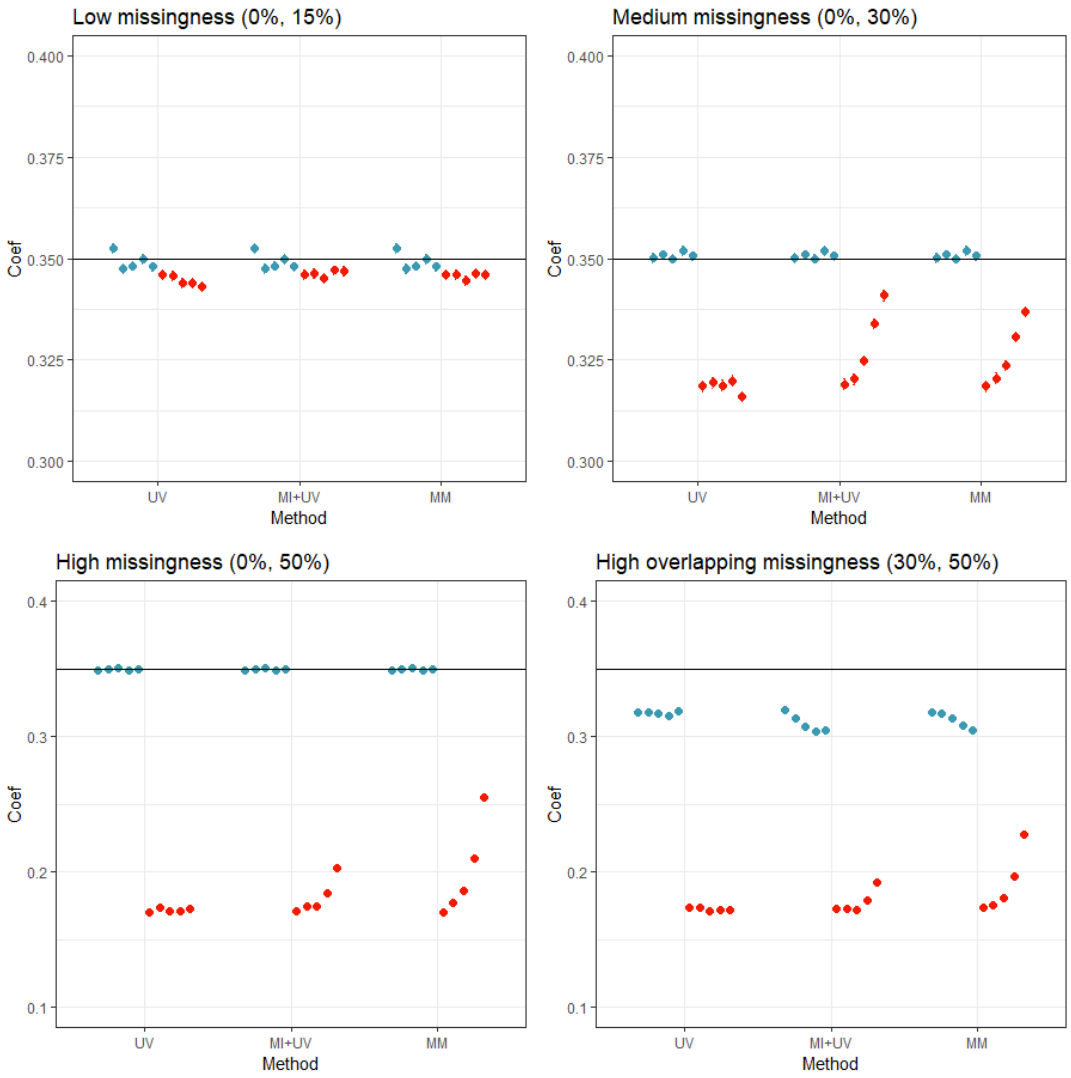


Figure 6.2 Bias in estimating intervention effect when simulating two binary outcomes and data are MNAR.

The blue dots represent the average of the estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 1. The red dots represent the average of estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 2. The five dots (of either colour) clustered together represents different correlation between the outcomes from 0 (left) to 0.8 (right) in increments of 0.2. Each graph corresponds to a different level of missing data as indicated. The true intervention effect is represented by the black horizontal line.

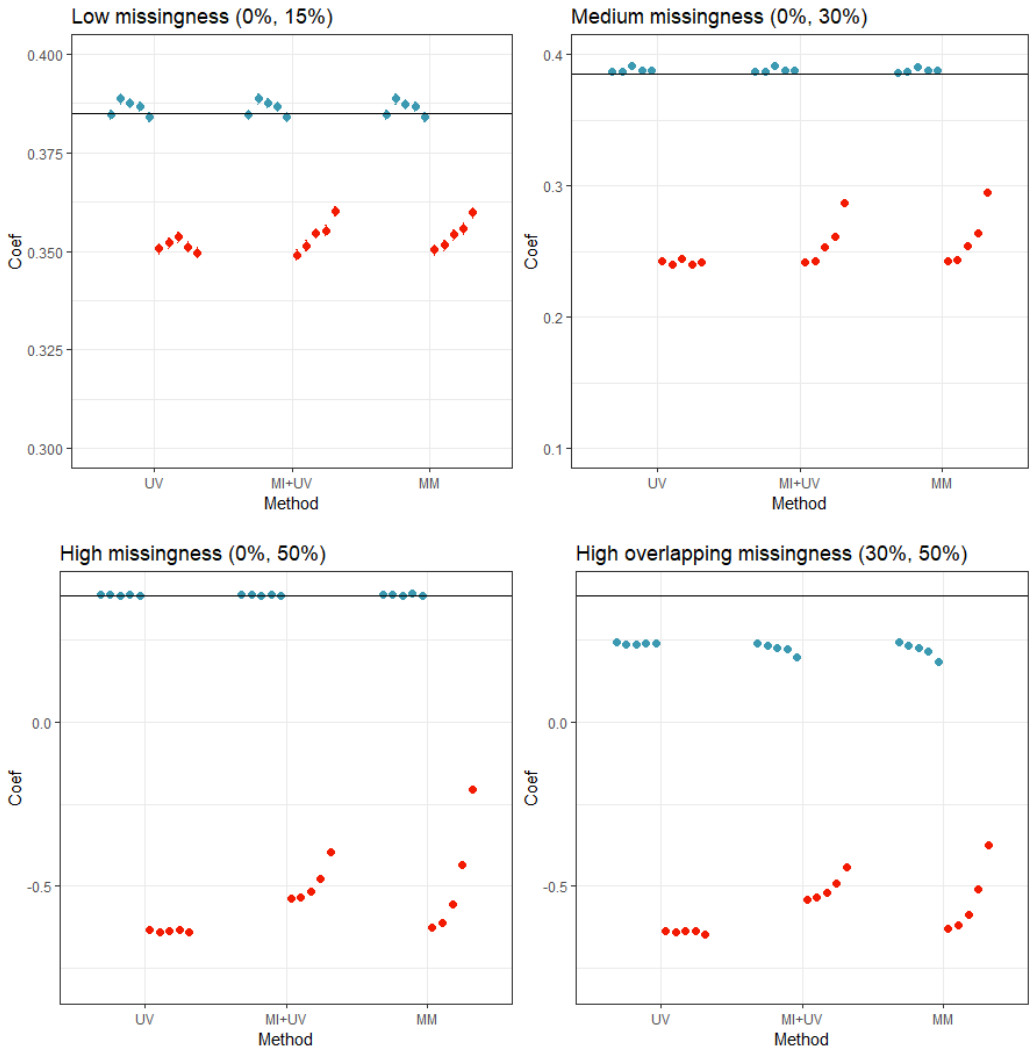


Figure 6.3 Bias in estimating intervention effect when simulating two 'mixed' outcomes and data are MNAR.

The blue dots represent the average of the estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 1. The red dots represent the average of estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 2. The five dots (of either colour) clustered together represents different correlation between the outcomes from 0 (left) to 0.8 (right) in increments of 0.2. Each graph represents a different level of missing data as indicated. The true intervention effect is represented by the black horizontal line.

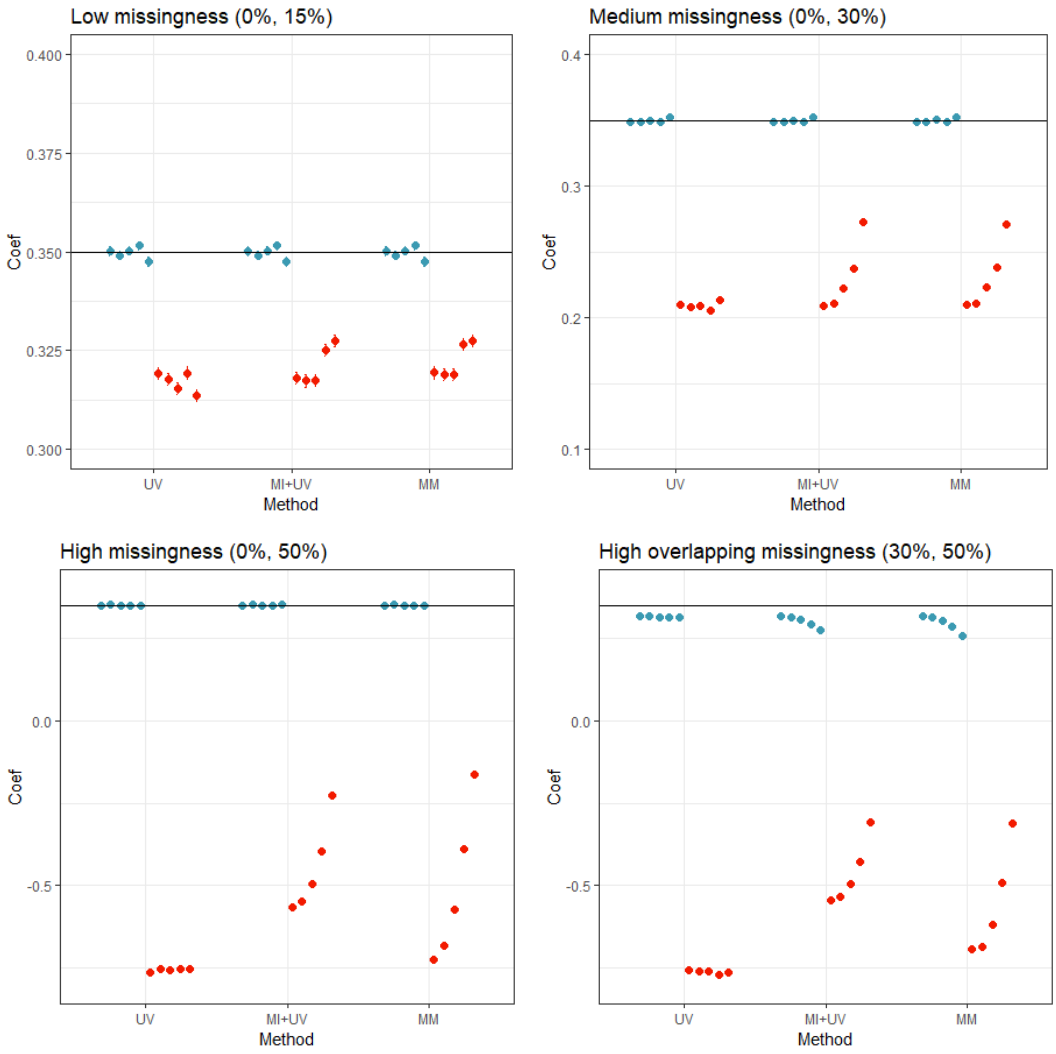


Figure 6.4 Bias in estimating intervention effect when simulating four continuous outcomes and data are MNAR.

The four colours each represent the average of the estimates treatment effects for the outcomes. The five dots (of each colour) clustered together represents different correlation between the outcome from 0 (left) to 0.8 (right) in increments of 0.2. Each graph represents a different level of missing data as indicated.. The bottom graph has high level of overlapping missing data. The true intervention effect is represented by the black horizontal line.

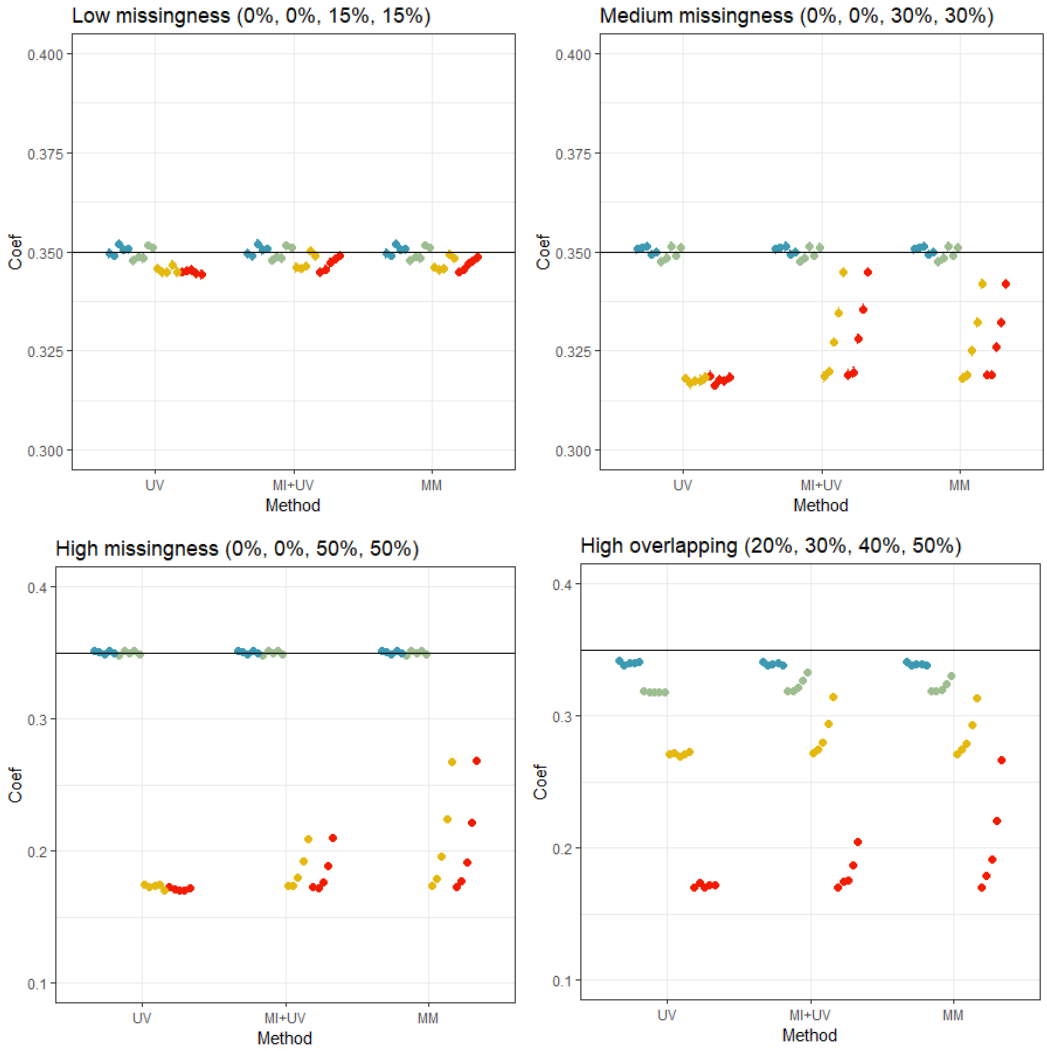
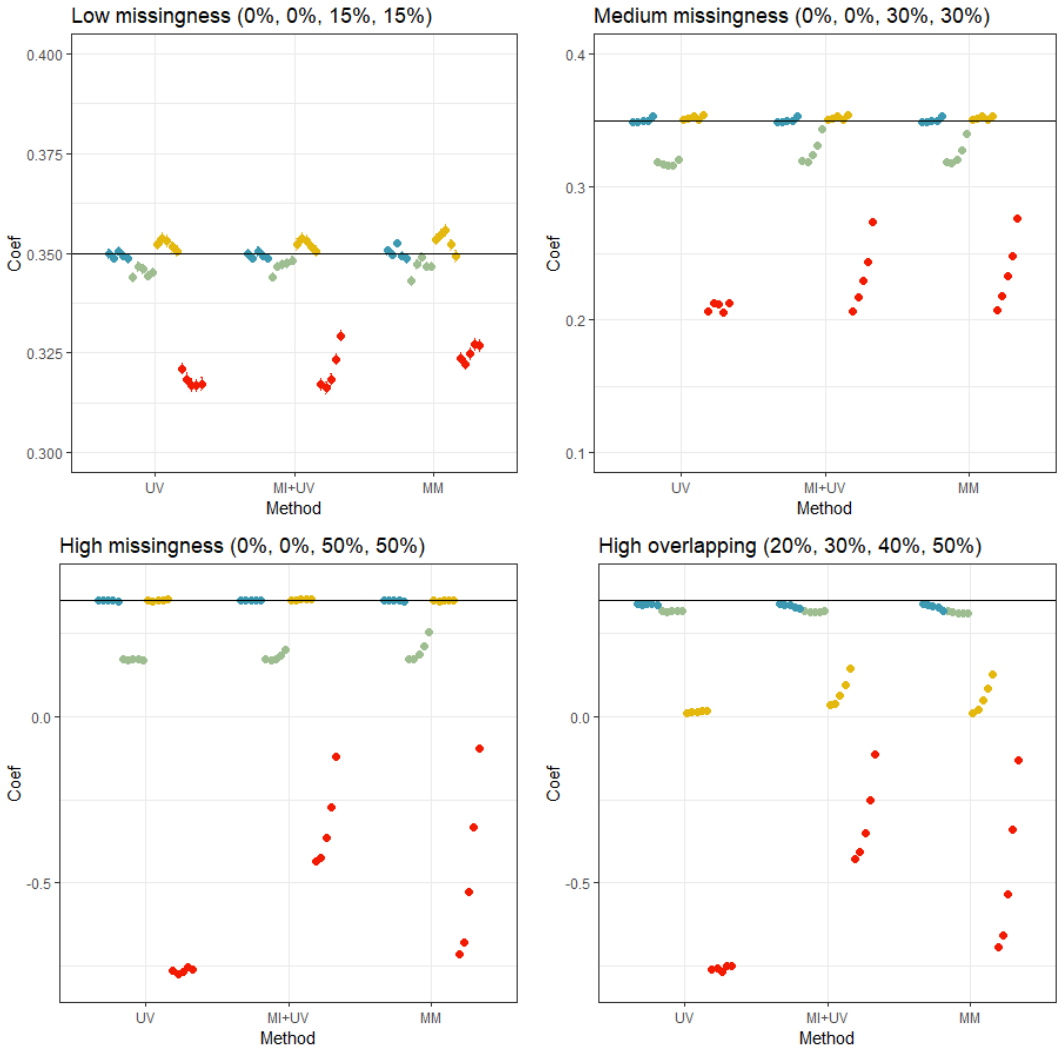


Figure 6.5 Bias in estimating intervention effect when simulating two continuous and two binary ('mixed') outcomes and data are MNAR.

The four colours each represent the average of the estimated treatment effects for the a outcomes. The five dots (of each colour) clustered together represents different correlation between the outcome from 0 (left) to 0.8 (right) in increments of 0.2. Each graph corresponds to a different level of missing data. The true intervention effect is represented by the black horizontal line.



6.4 Discussion

In this section, I investigated the bias in the estimated effects of the intervention when using the multivariate multilevel (MM) model compared to the bias observed when analysing the outcomes separately. The investigation included multiple imputation to handle any missing outcome values that are MNAR.

I have shown that when the MI+UV approach and MM model are used there is no reduction in bias in the estimated intervention effects if there is no correlation between the outcomes. However, there was a reduction in bias in the estimated intervention effects using both the MI+UV and MM methods when the outcomes are strongly correlated and in the presence of high levels of missing data. There was a notable reduction in bias when the correlation exceeds 0.4. The MM model appeared to outperform the MI+UV approach in the more extreme cases of high levels of missing data. However, as expected neither approach was able to remove the bias entirely. As a consequence, any inferences and conclusions made within the trial setting should be confirmed with sensitivity analyses under the alternative assumption that the missing data are MNAR. One approach would be to use MI to impute the missing values under the MAR assumption and to modify the MAR-imputed data to reflect a range of plausible scenarios. This can be obtained by multiplying the imputed values by a constant or by adding a constant to the imputed values. The adjusted results could be analysed by following the standard MI method, by analysing each imputed dataset and then combining the results.

Further reductions in terms of bias of the estimated intervention effects may have been observed if additional covariates had been included in the multiple imputation model that had taken account of the reason for missingness. By adding more variables into the model, it makes the MAR assumption more likely.

6.5 Conclusions

I found that the bias in the estimated effects of the intervention may be reduced by using either multiple imputation prior to analysing the outcomes separately or the multivariate multilevel model rather than analysing complete cases only. In the majority of the scenarios, very similar average estimates of the effects of the intervention were obtained when implementing the multiple imputation approach and multivariate multilevel model resulting in similar levels of bias of the estimated intervention effects.

Chapter 7

Evaluation of methods to jointly analyse continuous outcomes and survival outcomes

7.1 Introduction

In previous chapters, the focus has been on multiple primary outcomes that are either all continuous, binary or a combination of the two. However, clinicians may also be interested in time-to-event (survival) outcomes, such as time until death, drug relapse, or discharge from hospital. A time-to-event and a continuous outcome may be specified together as the primary outcomes in a trial.

The Contingency intervention for the Reduction of Cannabis use in Early psychosis (CiRCLE) trial is an example of when time-to-event and continuous outcomes were collected (Johnson et al., 2016). In the CiRCLE trial, the time to relapse was the single primary outcome, but they also measured positive symptom severity (using Positive and Negative Syndrome Scale, PANSS), which was also considered to be a key outcome (Kay et al., 1987). The outcomes were measured at three months and eighteen-months after the baseline measurement. It has been shown that cannabis use is associated with increased psychotic symptoms (Seddon et al., 2015). Consequently, it is expected that the time to relapse and PANSS scores (which measures psychotic symptoms) have a strong association.

Often when measuring a continuous outcome alongside a time-to-event outcome in trials, the continuous outcome is collected at multiple time points (longitudinally) over the follow up period. Joint models can be used to link the time-to-event outcome (relapse) with the continuous outcome (PANSS) to provide more accurate estimates of the effects of the intervention (Lawrence Gould et al., 2015).

In recent years, studies have investigated joint modelling of time-to-event outcomes with longitudinal continuous outcomes. Lawrence Gould et al. (2015) provide a summary of currently available joint models with an emphasis on Bayesian approaches. Ibrahim et al. (2010) also review joint modelling methods but focus on the bias observed in the time-to-event outcome when implementing joint models. In contrast, other studies have focused on

estimating the magnitude of the association between the time-to-event and continuous outcomes (Wang et al., 2012, Hatfield et al., 2011). By estimating the magnitude of the association between the outcomes, investigators can ascertain whether the survival time is associated with the continuous outcome for an individual. It has also been shown that joint models provide more efficient estimates of the effects of the intervention on the time-to-event and the continuous outcome, compared to analysing the two outcomes separately (Ibrahim et al., 2010, Wang et al., 2012). Several approaches for joint modelling have been suggested, however, it is not clear which of these models perform best in terms of bias and efficiency for both the time-to-event outcome and the continuous outcome.

7.2 Aim

The overarching aim of this chapter is to evaluate joint models to simultaneously analyse time-to-event and continuous outcomes. The specific objectives are: to review the existing methods that may be used to jointly model time-to-event and continuous outcomes; and to evaluate the performance of joint models in terms of bias and efficiency for the estimated effect of the intervention for both the time-to-event and continuous outcomes. The results will be compared to those obtained when analysing the outcomes separately.

7.3 Methods to analyse time-to-event and longitudinal outcomes

In this section, I describe methods that have been proposed to jointly analyse time-to-event and longitudinal continuous outcomes in RCTs. Joint models generally consist of two sub-models: one for the time-to-event outcome and one for the longitudinal continuous outcome.

The longitudinal continuous observations are usually modelled using a linear mixed model (Lawrence Gould et al., 2015)

$$Y_{ik} = \beta_{0i} + \beta_1 x_{ik} + z_{ik} u_i + \epsilon_{ik}.$$

Here, the Y_{ik} are the observed outcome values for participant i at time k. The β_{0i} are intercepts that vary for each participant. The x_i are binary variables that represent whether the participant is in the intervention group $(x_i = 1)$ or not $(x_i = 0)$ and β_1 is the coefficient for the effect of the intervention. The u_i are the random effects, which are normally distributed with a mean of zero and an unknown covariance parameter, that correspond to

the time-varying random effects z_{ik}. Lastly, ϵ_{ik} is the random error term, which is also normally distributed with a mean of zero and an unknown covariance parameter. The model can easily be extended to also include additional covariates, for example, baseline assessments. Equivalently, the linear mixed model may be written as

$$Y_i(t) = m_i(t) + \epsilon_i(t),$$

where the complete 'true' unknown patient-specific longitudinal trajectory be denoted by m_{ik}.

In the following sections, I discuss approaches to model a time-to-event outcome that is associated with the longitudinal continuous outcome.

Analysing the outcomes separately

The time-to-event outcome may be modelled separately to the longitudinal continuous outcome without any link between the two models. One model that is widely used to investigate the effect of the intervention on time-to-event outcomes is the Cox regression model (Cox, 1972). It models the hazard function denoted by $h(t)$, which is the risk of dying at time t given the individual is alive at time t. It is defined as follows

$$h(t) = h_0(t)\exp(\phi x_i).$$

Here, $h_0(t)$ is the baseline hazard at time t, x_i represents whether the participant is in the intervention group $(x_i = 1)$ or not $(x_i = 0)$ and ϕ is the associated log hazard ratio. The model could be extended to include additional covariates, for example, baseline assessments as before.

A key assumption for the Cox regression model is that the hazard functions for the two groups of participants should be proportional at all time points. Due to this, it is referred to as a proportional hazards model. One of the reasons for the popularity of the Cox proportional hazards model is that no assumptions are required about the underlying probability distribution of the outcome data (Cox, 1972) (Henderson et al., 2000). However, bias in the estimated treatment effect may occur and an unspecified baseline hazard may lead to an underestimation of standard errors of estimated treatment effect (Hsieh et al., 2006).

Other models can be used to analyse the time-to-event outcome, such as the exponential, Weibull or Gompertz distributions. These survival models do, however, make assumptions about the underlying probability distribution of the outcome data.

Shared parameters model

The idea behind the shared parameter joint model is to link the longitudinal and survival models via shared parameters. Let $M_i(t) = \{m_i(s); 0 \leq s \leq t\}$ denote the corresponding true, but unknown, longitudinal profile up to the survival time t. Then the proportional hazards model is defined as

$$h(t|M_i(t), x_i) = h_0(t) \exp\big(\phi x_i + \alpha m_i(t)\big)$$

where $h_0(t)$ is the baseline hazard function at time t. x_i represents if the participant is in the intervention group or not, ϕ is the associated log hazard ratio and α is the *association* parameter. $\exp(\alpha)$ is the hazard ratio for a unit increase in $m_i(t)$ at time t. By including the true unobserved trajectory function $m_i(t)$ in the linear predictor of the proportional hazards model, it is possible to link the longitudinal model and the proportional hazards model by a joint model. This joint model assumes that the association is based on the current value of the longitudinal response at time t. Once again, this model could be extended to include additional covariates, as necessary.

Joint random effects models

An alternative method is to use only the random effects in the linear predictor of the survival model. The random effects u_i are taken from the longitudinal model and are time-independent. Using joint random effects, the hazards model

$$h(t|M_i(t), x_i) = h_0(t) \exp(\phi x_i + \alpha u_i),$$

includes both the population level mean of the random effect, plus a subject specific deviation (Henderson et al., 2000).

When using the joint random effect model, the time-to-event data may be modelled using a Cox proportional hazards regression model with time-varying covariates. Although, as discussed earlier, other survival models may also be used. The longitudinal outcome is usually modelled using a linear mixed effects model. The association is captured by the joint random effects.

Correlated random effects model

Longitudinal outcomes and time-to-event data outcomes may also be jointly modelled via correlated random effects models (Philipson et al., 2012). One approach to analyse the time-

to-event outcome is to use a Cox proportional hazards model with a log-Gaussian frailty. The longitudinal model and the survival model are linked by allowing the Gaussian random effects of the linear model to be correlated with the frailty term of the Cox proportional hazards model. The model specifies latent vectors u_i and v_i that follow a zero-mean multivariate distribution, which are drawn independently for each participant. Given u_i and v_i, the longitudinal model is a linear mixed model as previously described. The hazards model is

$$h_i(t) = h_0(t)\exp(\phi x_i + \gamma_{ik} v_i)$$

where $h_i(t)$ is the hazard for subject i at time t, the x_i are the binary variables that indicate whether the participant is in the intervention group ($x_i = 1$) or not ($x_i = 0$), ϕ is the associated log hazard ratio and the γ_{ik} are time-varying explanatory variables. The models are linked via the multivariate distribution of the random effects u_i and v_i.

7.4 Software for the joint modelling of time-to-event and longitudinal outcomes

In this section, I discuss mainstream statistical packages to jointly model longitudinal and time-to-event outcomes. For the statistical software environment R, a variety of routines are available including joineR (Philipson et al., 2012), jointModel (Rizopoulos, 2010), frailtyPack (Rondeau et al., 2012), joineRML. The methodology for each routine is described below.

The joineR package implements the correlated random effects model. The JointModel package implements the shared parameter model (Rizopoulos, 2010). Leaving the baseline hazard function unspecified in JointModel leads to an underestimation of the standard errors of the parameter estimates (Hsieh et al., 2006). Consequently, even though an unspecified baseline hazard function is one of the options in the package, it is not recommended (Yuen and Mackinnon, 2016). Other distribution options, such as Weibull or Gamma or more flexible models based on spline-based approaches should be used instead. The joint model implemented in the frailtyPack package estimates simultaneously the longitudinal and survival processes using the relationship via random effects (Rondeau et al., 2012, Król et al., 2017). This package can also be used to jointly model longitudinal outcomes; recurrent events, for example hospital admissions; and terminal events, for example, death.

I have focused on the packages available in R. However, the package stjm is available in Stata (Crowther et al., 2013) and the package JMfit (Zhang et al., 2016a) is available in SAS.

Most of the packages mentioned only allow for one longitudinal outcome. Given that investigators should try and limit the number of primary outcomes, this is unlikely to be an issue in most trials. It should be noted, however, that the R package joineRML can be used if multiple longitudinal outcomes are required.

7.5 Simulation study

A simulation study was used to compare the performance of the joint models in terms of the bias and efficiency of the estimated effect of the intervention for both the time-to-event and continuous outcomes. The results will be compared to those obtained when analysing the outcomes separately.

Scenarios were simulated by varying the strength of the association between the longitudinal continuous outcome and the time-to-event outcome and the level of missing data for the continuous outcome. Details of these scenarios are provided in Table 7.1.

The simulated datasets contain a single continuous longitudinal and a single time-to-event outcome, which may be correlated. I simulated the data using a joint model that only shared the random effects. The random effects had a mean of zero. The longitudinal model contained a fixed intercept, time covariate and a binary intervention assignment covariate. The survival model was adjusted by only the binary intervention assignment covariate. It was assumed that the event was terminal and therefore no longitudinal information was recorded for an individual after their survival time. To generate the data the "simjointmeta" package in R was used.

Table 7.1 Scenarios simulated to evaluate methods which may be used to analyse a time-to-event outcome and a continuous outcome

Variable	Values
Association parameter	0, 0.5, 1, 1.5
Percentage of missing data values	1) No additional missing data 2) 25% of the continuous outcome was set to missing using a missing completely at random mechanism ('additional missing data').

I set the sample size to 560 participants, as this approximately the number in the motivating CiRCLE trial dataset, with an equal number of participants being allocated to each of the two randomised groups. For the longitudinal continuous outcome, I specified three follow up time points and a standardised intervention effect of 0.25 at both follow up time points. When assuming a significance level of 5%, the longitudinal outcome was individually powered at 84% when the association parameter between the models was zero. For the time-to-event outcome, a standardised intervention effect of 0.5 was chosen. The time-to-event outcome was censored such that there was approximately 50% censoring.

The bias associated with the estimated intervention effects was of primary interest. I also calculated: the FWER, the marginal power for each outcome, the overall disjunctive power, the coverage of the estimated intervention effects, the empirical standard error (EmpSE) and the mean square error of the estimated intervention effects and the Monte Carlo standard error (MCSE) of bias for the estimated intervention effects.

To account for multiple outcomes when calculating the FWER and the disjunctive power, the Bonferroni method was used. For each scenario, I ran 2500 simulations.

The following methods were used to analyse the data:

1) Univariate models (UV). For the longitudinal continuous outcome, a linear mixed model was implemented. For the time-to-event outcome a Cox proportional hazards model was implemented.
2) Correlated random effects models (using the R package JoineR)
3) Shared parameter estimates (using the R package JointModel)
4) Shared random effects (using the R package FrailtyPack)

When using the JointModel package, I specified the baseline hazard as a piecewise-constant function, meaning that the baseline hazard was specified to have different constant values within different time intervals. Additionally, the JointModel package offers two options for numerical integration: the standard Gauss-Hermite rule and the pseudo-adaptive Gauss-Hermite rule. It has been shown that the latter can be more effective in that typically fewer quadrature points are required to obtain an approximation error of the same magnitude and computational burden is reduced (Rizopoulos, 2010). Consequently, the latter was used in the analyses using the JointModel package. When implementing the models, the default settings were used for all other options.

7.6 Results

The estimated intervention effect together with the estimated bias of the estimated intervention effects are shown in Figure 7.1 and Figure 7.2. The figures show that when the outcomes are analysed separately, the estimator for the time-to-event outcome can be biased. When the association parameter was set to zero, no bias was observed, however, bias was observed when the association parameter was set to 0.5. The bias further increased when the association parameter was increased. When the association parameter was set to 1.5, the univariate Cox proportional hazards model significantly underestimated the intervention effect. The estimated bias in the estimated intervention effect for the survival model is reduced when implementing the joint models. When the association parameter was set to 1.5, the correlated random effects model (implemented using JoineR) provided the least biased estimates. The estimator of the intervention effect for the longitudinal model is unbiased when the association parameter was set to 0 or 0.5, as found for each of the methods. When the association parameter was set to 1.5, a small downward bias was observed for the univariate approach. The correlated random effects model (implemented using JoineR) and the shared random effects model (implemented using FrailtyPack) approach produced the least biased estimates. Similar results were observed for when 25% of the continuous outcome was set to missing (this scenario has been referred to as 'additional missing data'). The under-coverage of the univariate time-to-event outcome reflects the downwards bias described above.

The joint models increase the empirical standard error compared to analysing the outcomes separately. There were noticeable differences in the empirical standard errors when the association parameter was set to 1.5 (Time-to-event outcome: Univariate = 0.116; JoineR = 0.159; Fatality Pack = 0.164; JointModel = 0.152). When the association parameter was set to 0 and 0.5, the mean square error observed was comparable for each of the methods. When the association parameter was set to 1.5, the joint random effect models (implemented using JoineR) and the model that utilised shared parameters between the longitudinal model and the survival model (implemented using JointModel) performed best in terms of mean square error.

The FWER was controlled at around 0.05 (ranging from 0.041 to 0.057) when there was no additional missing data in the longitudinal outcome. When there was missing data, the FWER slightly increased for the shared random effects model (implemented using the frailty pack) ranging from 0.052 to 0.063 depending on the magnitude of the association parameter.

The marginal power for both outcomes was reduced as the association parameter increased. When the association parameter was set to zero, the longitudinal outcome had 84%-85% power and the time-to-event outcome had 99% power. The higher marginal power in the time-to-event outcome is likely to have dominated the disjunctive power results.

Figure 7.1 Bias in estimating the intervention effects when simulating one time-to-event and one continuous outcome and no additional missing data in the continuous outcome.

The red dots represent the average of the estimated intervention effect ($\hat{\beta}$) for the time-to-event outcome, with error bars representing $\pm 1.96 \times \text{Monte CarloSE(Bias)}$. The light blue dots represent the average of the estimated intervention effect ($\hat{\beta}$) for the continuous longitudinal outcome, with error bars representing $\pm 1.96 \times \text{Monte CarloSE(Bias)}$. The association parameter varied across the graphs. This is described by the "association" in the headings for each of the graphs. The true intervention effects on both outcomes are represented by the grey horizontal line.

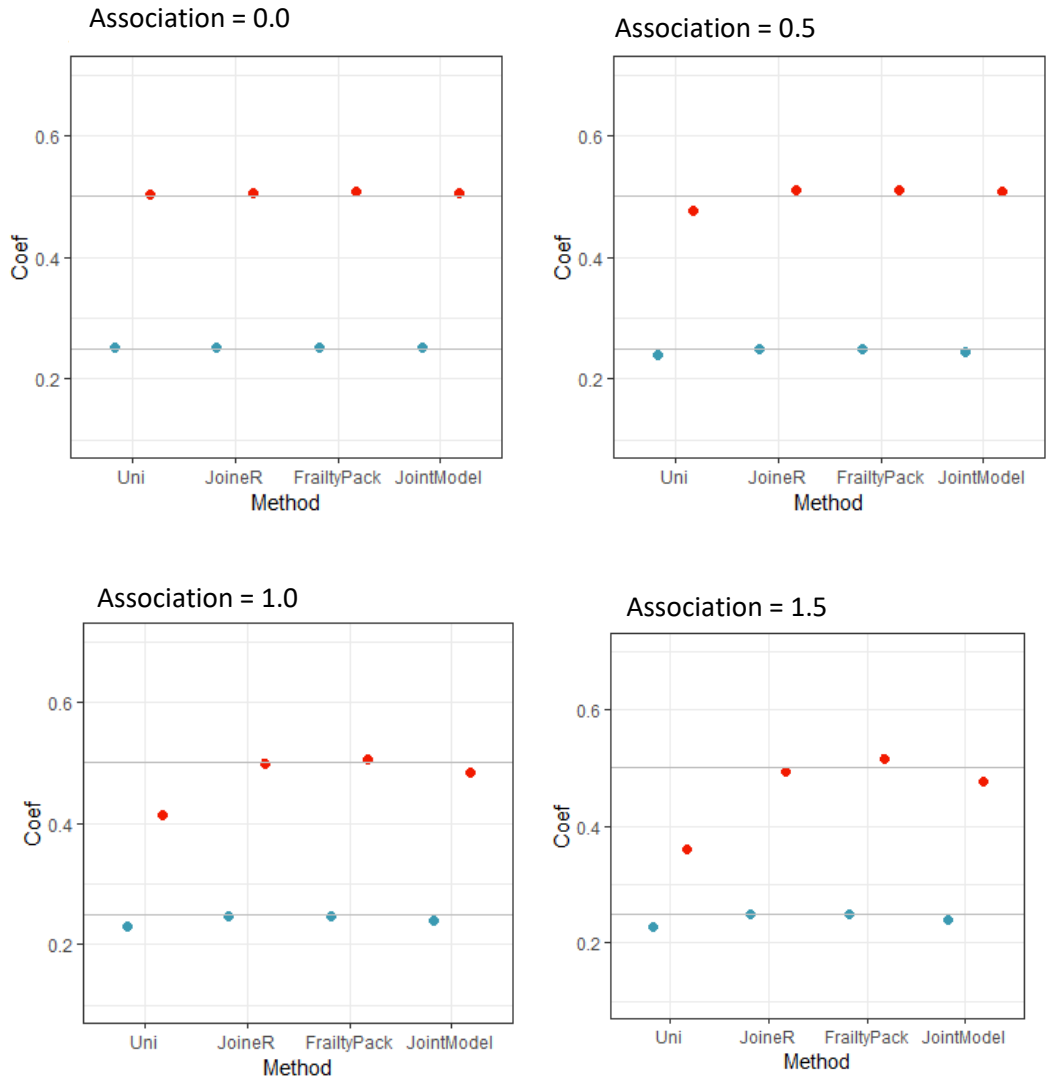


Figure 7.2 Bias in estimating intervention effects when simulating one time-to-event and one continuous outcome with additional missing data in the continuous outcome.

The red dots represent the average of estimated intervention effect ($\hat{\beta}$) for the time-to-event outcome, with error bars representing $\pm 1.96 \times$ Monte CarloSE(Bias). The light blue dots represent the average of the estimated intervention effect ($\hat{\beta}$) for the continuous longitudinal outcome, with error bars representing $\pm 1.96 \times$ Monte CarloSE(Bias). The association parameter varied across the graphs. This is described by the "association" in the headings for each of the graphs. The true intervention effects on both outcomes are represented by the grey horizontal line.

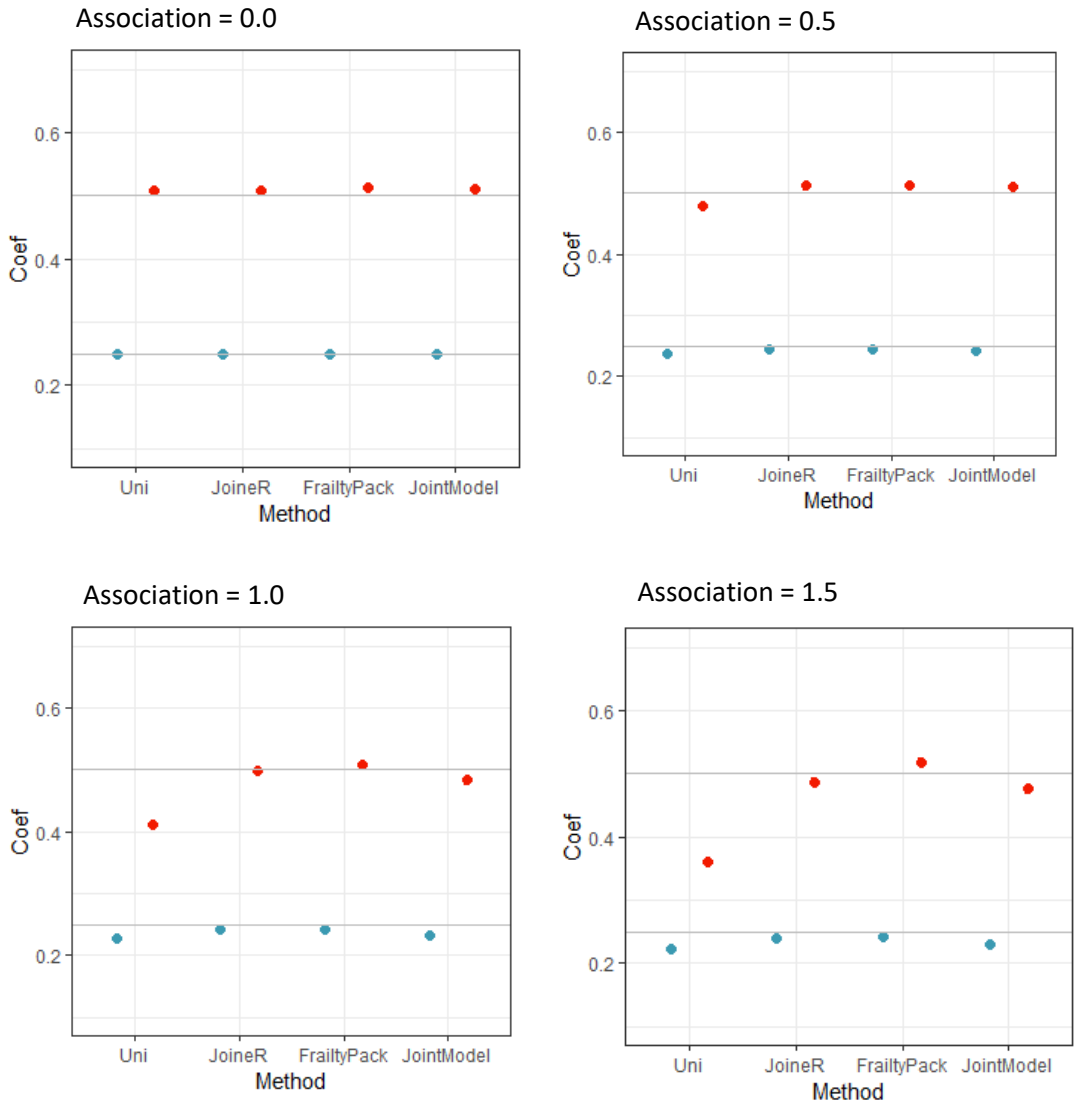


Table 7.2 Coverage of the estimated intervention effects obtained when evaluating one time-to-event outcome and one longitudinal outcome.

The univariate model is compared to three joint models. The scenarios evaluated vary by the type of missing data and the magnitude of the association parameter. The nominal coverage probability is 95.0%

Additional missingness for continuous outcome ↓	Association between outcomes ↓	**Continuous longitudinal outcome**				**Time-to-event outcome**			
	Method →	Uni	JoineR	Frailty Pack	Joint Model	Uni	JoineR	Frailty Pack	Joint Model
None	0	94.8	94.4	94.7	94.7	94.1	94.1	93.7	94.0
	0.5	95.2	94.8	95.2	95.8	93.4	94.0	94.0	93.7
	1	94.6	95.1	95.6	95.4	88.0	95.2	95.2	93.8
	1.5	95.0	96.0	96.2	96.0	76.2	95.5	95.7	93.2
MCAR 25%	0	94.9	94.5	94.6	94.7	94.9	94.8	94.8	95.0
	0.5	93.2	94.2	94.0	94.1	93.6	95.0	94.4	94.6
	1	94.6	94.2	94.8	94.8	90.4	95.0	95.6	94.6
	1.5	94.4	94.8	94.8	95.4	76.8	96.4	96.0	94.8

Table 7.3 Empirical standard error of the estimated intervention effects obtained when evaluating one continuous outcome and one time-to-event outcome

The univariate model is compared to three joint models. The scenarios evaluated vary by the type of missing data and the magnitude of the association parameter.

Additional missingness for continuous outcome ↓	Association between outcomes ↓	**Continuous longitudinal outcome**				**Time-to-event outcome**			
	Method →	Uni	JoineR	Frailty Pack	Joint Model	Uni	JoineR	Frailty Pack	Joint Model
None	0	0.083	0.083	0.083	0.083	0.119	0.12	0.12	0.119
	0.5	0.083	0.084	0.084	0.084	0.118	0.126	0.126	0.125
	1	0.081	0.082	0.082	0.082	0.116	0.139	0.14	0.134
	1.5	0.081	0.083	0.083	0.082	0.116	0.159	0.164	0.152
MCAR 25%	0	0.094	0.094	0.094	0.094	0.12	0.121	0.121	0.121
	0.5	0.093	0.093	0.093	0.093	0.124	0.131	0.132	0.130
	1	0.090	0.091	0.091	0.090	0.114	0.141	0.143	0.136
	1.5	0.089	0.091	0.090	0.089	0.122	0.166	0.180	0.162

Table 7.4 The Monte Carlo standard errors of the estimated intervention effects obtained when evaluating one continuous outcome and one time-to-event outcome.

The univariate model is compared to three joint models. The scenarios evaluated vary by the type of missing data and the magnitude of the association parameter.

Additional missingness for continuous outcome ↓	Association between outcomes ↓	**Continuous longitudinal outcome**				**Time-to-event outcome**			
	Method →	Uni	Joine R	Frailty Pack	Joint Model	Uni	Joine R	Frailty Pack	Joint Model
None	0	0.007	0.007	0.007	0.007	0.014	0.014	0.014	0.014
	0.5	0.007	0.007	0.007	0.007	0.014	0.016	0.016	0.016
	1	0.007	0.007	0.007	0.007	0.021	0.019	0.02	0.018
	1.5	0.007	0.007	0.007	0.007	0.033	0.025	0.027	0.024
MCAR 25%	0	0.009	0.009	0.009	0.009	0.014	0.015	0.015	0.015
	0.5	0.009	0.009	0.009	0.009	0.022	0.021	0.021	0.02
	1	0.009	0.008	0.008	0.008	0.020	0.020	0.020	0.019
	1.5	0.009	0.008	0.008	0.008	0.035	0.028	0.033	0.027

Table 7.5 The familywise error rate obtained when evaluating one continuous and one time-to-event outcome

The univariate model is compared to three joint models. The scenarios evaluated vary by the type of missing data and the magnitude of the association parameter. The Bonferroni method was used to account for the multiplicity.

Additional missingness for continuous outcome ↓	Association between outcomes ↓				
	Method →	Uni	Joine R	Frailty Pack	Joint Model
None	0	0.041	0.045	0.042	0.040
	0.5	0.052	0.057	0.050	0.050
	1	0.050	0.052	0.055	0.049
	1.5	0.048	0.047	0.052	0.045
MCAR 25%	0	0.055	0.056	0.063	0.055
	0.5	0.050	0.052	0.057	0.05
	1	0.055	0.056	0.063	0.055
	1.5	0.040	0.043	0.052	0.038

Table 7.6 The marginal power obtained for each of the outcomes when evaluating one continuous outcome and one time-to-event outcome.

The univariate model is compared to three joint models. The scenarios evaluated vary by the type of missing data and the magnitude of the association parameter. No adjustment has been made to account for the multiplicity.

Additional missingness for continuous outcome ↓	Association between outcomes ↓	**Continuous longitudinal outcome**				**Time-to-event outcome**			
	Method →	Uni	Joine R	Frailty Pack	Joint Model	Uni	Joine R	Frailty Pack	Joint Model
None	0	0.844	0.847	0.848	0.847	0.992	0.99	0.992	0.991
	0.5	0.812	0.845	0.842	0.827	0.987	0.989	0.989	0.992
	1	0.801	0.841	0.845	0.82	0.948	0.956	0.961	0.961
	1.5	0.785	0.826	0.835	0.807	0.88	0.881	0.895	0.894
MCAR 25%	0	0.760	0.752	0.761	0.76	0.987	0.986	0.988	0.987
	0.5	0.700	0.733	0.749	0.714	0.931	0.937	0.941	0.946
	1	0.690	0.744	0.756	0.712	0.938	0.944	0.952	0.954
	1.5	0.706	0.739	0.766	0.720	0.826	0.835	0.850	0.862

Table 7.7 The overall disjunctive power obtained when evaluating a continuous outcome and a time-to-event outcome

The univariate model is compared to three joint models. The scenarios evaluated vary by the type of missing data and the magnitude of the association parameter. The Bonferroni method has been used to account for the multiplicity.

Additional missingness for continuous outcome ↓	Association between outcomes ↓				
	Method →	Uni	Joine R	Frailty Pack	Joint Model
None	0	0.996	0.991	0.996	0.996
	0.5	0.985	0.987	0.990	0.990
	1	0.970	0.970	0.985	0.981
	1.5	0.934	0.947	0.967	0.960
MCAR 25%	0	0.993	0.988	0.993	0.993
	0.5	0.978	0.978	0.978	0.985
	1	0.951	0.953	0.963	0.962
	1.5	0.882	0.897	0.923	0.912

7.7 Discussion

In this chapter, I performed a simulation study to investigate the differences in bias obtained when using three different joint modelling approaches to jointly analyse a time-to-event outcome and a continuous outcome in a trial. The results were compared to those obtained using separate univariate models. Additionally, I quantified the differences in disjunctive power and FWER achieved using the different methods.

The routines have been implemented in the statistical software R to enable the joint modelling to be "user-friendly" and more readily accessible. The time to estimate the different models varied considerably. The shared parameter model took several minutes longer to implement than the other methods. This is because it used bootstrapping to calculate the confidence intervals. In practice, when analysing a single dataset, bootstrapping only needs to be performed once, which in usual cases would not significantly delay the analysis.

The interpretation of the estimated intervention effect for the time-to-event outcome varied across the different R packages. When implementing the models with shared random effects (using the joineR package), the intervention effects are specific to each of the models. As a result, the estimates are easy to interpret. In contrast, when using the shared parameter model (implemented using the JointModel package) the overall intervention effect is decomposed into two parts: the direct and indirect effects. The direct component stems from parameters being included in the survival model as fixed effects. The indirect components link the survival model to the estimated coefficients calculated by the longitudinal model. The two intervention effects may be combined to provide the overall intervention effect. The overall effect is the sum of the direct component plus the product of the relevant association parameter and the indirect component (Ibrahim et al., 2010). The combined intervention effect is comparable with the intervention effects estimated using the other methods. Whilst it is easy to calculate the combined estimated intervention effect, it is not as straightforward to combine the variances. If the confidence intervals are required, then the other methods may be easier to implement.

The univariate Cox model underestimated the intervention effect, which resulted in a downward bias of the estimated intervention effects for the time-to-event outcome. The bias was reduced when using any of the joint models investigated. The FWER is maintained at an acceptable level around 0.05 when implementing the univariate model of the joint models via the JoineR routine or JointModel routine, however, when implementing

fraitlypack, the FWER is increased for some scenarios. The disjunctive power was increased when using the joint models compared to using the univariate models. The greatest benefit in terms of disjunctive power was observed when there was additional missing data in the continuous outcome and high association between the outcomes. Taking into account these simulation results and the interpretation of the estimated intervention effect discussed earlier I would recommend that the JoineR routine is used.

One feature of the joint models examined in this chapter is that the time-to-event outcome must be terminal. As a result of this assumption, the models do not consider any longitudinal measurement after the event of interest has occurred. However, there are scenarios in which the event is not terminal and longitudinal measurements continue to be collected after the event. For example in the CiRCLE trial described in Section 7.1, the event of interest was time to relapse, defined as admission to an acute mental health service. The longitudinal outcomes, which included positive psychotic symptoms, were measured at 3 months and 18 months regardless of if the participant had an event or not. In this example, 88 participants relapsed prior to the 18 month follow-up but still provided data at this time point. If this data was removed from the analysis, incorrect conclusions may be drawn regarding the effect of the intervention on the longitudinal outcomes.

Study extensions and limitations

The joint modelling framework may also be used to assess the effect of the longitudinal outcomes on the probability of the event occurring. This may be useful in psychiatry trials or palliative care trials where it is likely that the continuous outcome is associated with the time-to-event outcome.

When simulating the data, it was assumed that no longitudinal data was collected after the event of interest. This reflects the situation when the event is terminal, for example, if a participant dies. In this instance, no further longitudinal information is collected on the participant. In some trials, however, the event may not be terminal and longitudinal data may still be collected after the event or alternatively the event may be recurrent. For example, in psychiatry trials the event may be time until return to drug use. In this instance, after the event (return to drug use) information may still be collected. Different methodology may be required to analyse such data (Mazroui et al., 2012, Kim et al., 2012, Liu et al., 2004).

The frailtyPack routine described earlier can be used to implement a joint model of longitudinal continuous outcomes and recurrent events.

Additionally, I only considered frequentist methods for the joint modelling of time-to-event and longitudinal outcomes, however, Bayesian methods have been proposed by Faucett and Thomas (1996), Ibrahim et al. (2010) and Wang and Taylor (2001). These Bayesian methods also use a proportional hazards model for the survival model, however, a different longitudinal model is used. Bayesian methods may be modelled in R using JMBayes (Zhang et al., 2016a) which uses OpenBUGS or WinBUGS (Lawrence Gould et al., 2015).

7.8 Conclusions

Joint models can be used to link time-to-event outcomes with continuous outcomes and could provide better more accurate estimates of the effect of the intervention. The time-to-event and continuous outcomes may be analysed using a survival model and a longitudinal model, respectively, and these models can be linked. The bias in the estimators for the time-to-event outcome is reduced when using the joint models.

Chapter 8

Discussion, guidance and conclusions

The work in this thesis was primarily motivated by the difficulties raised by clinicians working on RCTs in the field of psychiatry. I have experienced first-hand, when designing trials,that clinicians can find it challenging to select a single primary outcome that encompasses all aspects of the health condition they are investigating. As a consequence, several outcomes are often chosen as the primary outcomes.

There are numerous approaches available to analyse RCTs with multiple outcomes. Previous work suggested that multivariate methods may produce more precise and accurate estimates of intervention effects when compared to univariate methods (Pituch et al., 2016, Snijders and Bosker, 2012). However, the extent of these benefits and the scenarios in which these benefits may be realised was not known. Additionally, there was a lack of guidance regarding which approach should be used to calculate the required sample size for an RCT when there are multiple primary outcomes. As a result, the overarching aim of this thesis was to understand which methodologies should be used to calculate the required sample size and to perform the analysis of an RCT that has multiple primary outcomes.

The specific objectives were to:

1. Investigate the frequency that multiple primary outcomes are recorded and analysed in published RCTs, and to investigate which methods are used for the sample size calculation and analysis of these trials.
2. Investigate which of the relevant adjustment methods should be used to control the FWER when analysing correlated primary outcomes.
3. Investigate which of the relevant methods should be used to analyse multiple primary outcomes and to determine the scenarios in which the methods should be used.

In the remainder of this chapter, I provide a brief summary of my findings, followed by recommendations on which methods should be used. Finally, I review possible areas for further research and present my overall conclusions.

8.1 Summary of thesis and findings

In Chapter 2, I provided a concise summary of the background and key concepts that are required when discussing multiple outcomes. I introduced the concept of 'alternative outcomes' and 'co-primary outcomes' and explained that the work in this thesis focuses on alternative outcomes. This is when the main clinical objective of a trial is formulated in terms of investigating the effect of the intervention on several primary outcomes, and the objective is met if at least one analysis produces statistically significant results. Following this, I provided the definitions of the familywise error rate (FWER) and disjunctive power. Both of these concepts need to be considered when selecting the method to analyse multiple primary outcomes. I then went on to discuss missing data theory. Almost all RCTs have outcomes that have missing values (Bell et al., 2014) and if the missing data are ignored or incorrectly handled then the conclusions drawn from the data could be incorrect (Carpenter and Kenward, 2007). One approach that I discussed to handle missing data was the use of multiple imputation.

I reviewed relevant methodologies that have been commonly used or recommended for use in the statistical analysis of multiple primary outcomes. I discussed some of the disadvantages of using global test statistics (including the necessity to have balanced data across the outcomes) and using factorisation modelling (including the lack of guidance on how to use this model for more than two outcomes). I observed that the multivariate multilevel model is rarely used in clinical trials. However, it is used in other areas of research and could be easily applied to clinical trials. I explained why the multivariate multilevel model and the latent variable model are my preferred methods amongst those discussed. Briefly, both these methods can handle continuous outcomes, binary outcomes or a combination of the two types. In addition, the number of observations does not need to be balanced across outcomes and the methods can easily be extended to handle more than two outcomes.

In Chapter 3, I reviewed RCTs that were published in high impact neurology and psychiatry journals. The review showed that multiple outcomes were commonly used but are often inadequately analysed. The majority of trials analysed outcomes separately without any adjustment for multiple comparisons. When adjustment methods were implemented, only the most basic methods were used. The Bonferroni approach was the most commonly used method, although the Holm, Hochberg and Šidák methods were also used. This review highlighted that multiple outcomes are being used in RCTs but guidance is needed regarding which methods should be used for the sample size calculation and analysis.

This leads on to Chapter 4 which investigated methodologies to control the FWER when analysing correlated multiple outcomes. One approach is to adjust the p-values from each statistical test used to investigate the effect of the intervention or alternatively the significance level used in the comparisons. When analysing multiple correlated outcomes, I recommend that either the Hommel or Hochberg method is used, assuming that the distributional assumptions are met. I highlighted that the sample size requirement to achieve the desired disjunctive power may be smaller than that required to achieve the desired marginal power. The choice between whether to specify a disjunctive or marginal power should depend on the clinical objective.

In Chapter 5, I evaluated multivariate methods for the analysis of multiple primary outcomes in clinical trials. I performed a simulation study to investigate the differences between the preferred multivariate methods and the standard univariate approach, with and without multiple imputation. The work focused on continuous outcomes, binary outcomes and a combination of the two types. My simulation results suggest that the power to detect an intervention effect may be increased by using multivariate multilevel models, rather than by analysing each outcome separately. However, I found that the power gains were small in all but the most extreme scenarios. The largest gains were observed when there was strong correlation between the outcomes and high levels of missing data. My findings are consistent with the results presented in Pituch et al. (2016) and Snijders and Bosker (2012). Additionally, the multivariate multilevel model does not require any prior imputation of missing data. The multivariate multilevel model is also flexible allowing both shared intervention effects and individual intervention effects to be estimated.

The work was extended in Chapter 6 to consider data which are missing not at random (MNAR). Under MNAR, parameter estimation from the observed data alone is typically biased. I investigated whether the multivariate multilevel model could reduce the bias in the estimated intervention effect when the missing data mechanism is MNAR. As expected, no reduction in terms of bias were made when there was no correlation between the outcomes. A notable reduction in bias for both the multivariate multilevel model and the multiple imputation approach occurred when there was moderate to high pairwise correlation between the outcomes.

In Chapter 7, I considered methods to analyse time-to-event outcomes alongside continuous outcomes. Joint models can be used to link time-to-event outcomes with continuous outcomes and these models may provide better estimates of the intervention effect

compared to analysing the outcomes separately. The simulation results showed that when the outcomes are analysed separately, parameter estimation for the time-to-event outcome is typically biased. The bias is reduced when using joint models. The largest reduction of bias in the estimates were observed when there was a strong association between the time-to-event and continuous outcomes.

8.2 Recommendations for reporting

In Chapter 3, I reviewed published randomised controlled trials. When extracting the information required for the review, I noted that key information was missing from some of the papers. For example, in some papers it was unclear which, if any, of the outcomes were deemed primary. As a result, I have made the following recommendations regarding the reporting of results of a clinical trial.

Once the authors have specified the methods for the sample size calculation and analysis, the protocol and journal article should be written in sufficient detail to ensure the reader is fully aware of the methods used. As advised by the current ICH guidelines, the trial objectives should be clearly stated (Phillips and Haudiquet, 2003). Furthermore, the authors should ensure that they have specified the primary and secondary outcomes, methods of measurements and time points of interest at the start of the trial (WHO, 2012). The documentation of the pre-specified outcomes is encouraged by the CONSORT checklist (Schulz et al., 2010). The sample size calculation should be based on all the primary outcomes (Chan et al., 2013). Authors should report the sample size calculation and state which of the outcomes are used in its calculation to ensure that the reader is aware of how the trial is powered.

With regards to multiplicity arising from multiple outcomes, CONSORT state that “authors should exercise special care when evaluating the results of trials with multiple comparisons” (Schulz et al., 2010). I recommend that the chosen method to maintain the FWER at the desired level is reported and justification for the choice provided. If the RCT is viewed as confirmatory, the ICH E9 guidelines state that any aspects of multiplicity should be identified in the protocol; adjustment should always be considered and the details of any adjusting method, or an explanation of why an adjustment is not thought to be necessary, should be set out in the analysis plan (Phillips and Haudiquet, 2003). The abstract should be clear, transparent, and sufficiently detailed (Hopewell et al., 2008), as explained in the CONSORT

statement. This is because readers often base their assessment of the trial on the information provided in the abstract. It is important that the abstract is an accurate record of the trial and is not in any way ambiguous or misleading.

There is no general consensus regarding the importance of secondary outcomes; they can be viewed as supportive evidence or as a basis for additional claims. If secondary outcomes are viewed as supportive evidence then statistical adjustments may not be required (Pocock, 1997). Appropriate caution should be exercised when interpreting their results. One option to ensure that secondary outcomes are given less emphasis would be to present estimates of the intervention effect for them with the corresponding confidence intervals, rather than the p-values. This would give information about the level of precision and whether the confidence level included a clinically important intervention effect. If the secondary outcomes are used for additional claims then multiplicity needs to be accounted for when analysing these outcomes too (Committee for Proprietary Medicinal Products, 2002). For example, further confirmatory statistical testing on secondary variables can be performed using a further hierarchical order for the secondary variables (Committee for Proprietary Medicinal Products, 2002).

8.3 Implementation of the recommended methods when analysing multiple outcomes

The sample size calculation for any clinical trial should reflect the clinical aims of the trial. If multiple primary outcomes are used then this should be reflected in the calculation. When calculating the required sample size for an RCT with multiple primary outcomes, I recommend that the Bonferroni method is used to account for multiplicity. To implement the Bonferroni method in the sample size calculation, any standard package may be used. For example, in Stata the power command may be used or in R the samplesize command may be used. The significance level would need to be adjusted according to the Bonferroni method.

During the analysis stage I recommend that either the Hochberg or Hommel method is used to control the FWER. To implement the Hochberg or Hommel p-value adjustment method, I recommend using the R package p.adjust. The Hochberg method may be implemented in Stata using the multproc command. There is no inbuilt function in Stata to use the Hommel method. However, if this method is desired, I would recommend that the user performs the analysis in their chosen software and copies the p-values into R for adjustment.

To analyse multiple continuous outcomes, multiple binary outcomes or a combination of the two, I recommend that the multivariate multilevel model is used. The multivariate multilevel model may be implemented using the statistical software package MLwiN (Rabash et al., 2009). MLwiN can be used via R and Stata using the R2MlwiN (Zhang et al., 2016b) and runMLwiN (Leckie and Charlton, 2013) packages, respectively. Details of how the multivariate multilevel model can be implemented in Stata and R are provided in Appendix 7.

To analyse a time-to-event outcome and a continuous outcome, I suggest that a joint model is used. In particular, I recommend that a correlated random effects model is used which may be implemented using the *JoineR* package in R.

8.4 Limitations and future work

Based on my review of published trials, I focused on the simpler methods to control the FWER. However, there are other more advanced methods available in the literature. For example, to control the FWER, graphical methods (Bretz et al., 2011, Bretz et al., 2009) or Dunnett's methods (Dunnett, 1955) may be used. The step-down Dunnett method and step-up Dunnett method require complicated, iterative procedures that have not been implemented in any statistical software (Blakesley et al., 2009). I therefore felt that the other methods were more relevant for the comparison. The graphical methods may be used to evaluate outcomes that have a pre-specified hierarchy (Bretz et al., 2009, Bretz et al., 2011). Graphical models may also be used when the analysis plan is complex due to splitting of the overall alpha among the outcomes. The graphical models are particularly useful if there is a desire to have a 'second chance' for an outcome that was not statistically significant at the initially assigned outcome-specific alpha. Outcomes that were not statistically significant initially receive 'pass-along' alpha from a different outcome that was statistically significant at the initially assigned alpha (Food and Drug Administration, 2017).

In the review of published RCTs, I observed that the majority of papers that analysed multiple primary outcomes specified two primary outcomes and very few papers used more than four outcomes. Consequently, in this thesis, I focused on providing recommendations for analysing two to four outcomes. However, in other areas of research such as genetic studies the number of outcomes being analysed maybe more than this. Further work is required to investigate which methods should be used when analysing a larger number of outcomes.

The focus of this work was on multiple primary outcomes, however, similar issues regarding multiplicity will also arise when analysing multiple intervention groups (Freidlin et al., 2008, Baron et al., 2013) or multiple patient populations (Brookes et al., 2001). Further work is required to see which methods would be best suited to analyse the data accounting for multiplicity in these scenarios.

I have not considered the Bayesian framework in this thesis. When a Bayesian framework is used, external evidence may be included in all aspects of an RCT, including the design, analysis and interpretation (Spiegelhalter et al., 2004). As a consequence, the Bayesian approach may be viewed as more efficient as it is able to make use of all available evidence rather than restricting the analysis to just the new data collected. Additionally, the Bayesian framework is valuable as it can provide a more flexible approach to the analysis that can be adapted to each trial (Spiegelhalter et al., 2004). In this work, I have often assumed that the outcomes are normally distributed, either directly or via a latent variable. In a Bayesian analysis, more complicated models can be used. This may be required when analysing health economic data, which often includes skewed cost data and utility values which lie between zero and one.

8.5 Conclusions

In this thesis, I addressed the need for a review and evaluation of how multiple outcomes are analysed in published randomised controlled trials. I also addressed the need for a comparison between univariate and relevant multivariate methods for the analysis of clinical trials. The multivariate multilevel model can be used to analyse clinical trials with multiple primary outcomes, which are correlated, to produce a more accurate estimate of the intervention effects.

References

AGUSTI, A. & VESTBO, J. 2011. Current controversies and future perspectives in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine,* 184**,** 507-13.

ALOSH, M., BRETZ, F. & HUQUE, M. 2014. Advanced multiplicity adjustment methods in clinical trials. *Statistics in medicine,* 33**,** 693-713.

BARON, G., PERRODEAU, E., BOUTRON, I. & RAVAUD, P. 2013. Reporting of analyses from randomized controlled trials with multiple arms: a systematic review. *BMC medicine,* 11**,** 84.

BEEKEN, R., CROKER, H., MORRIS, S., LEURENT, B., OMAR, R., NAZARETH, I. & WARDLE, J. 2012. Study protocol for the 10 Top Tips (10TT) Trial: Randomised controlled trial of habit-based advice for weight control in general practice. *BMC public health,* 12**,** 667.

BEEKEN, R., LEURENT, B., VICKERSTAFF, V., WILSON, R., CROKER, H., MORRIS, S., OMAR, R., NAZARETH, I. & WARDLE, J. 2017. A brief intervention for weight control based on habit-formation theory delivered through primary care: results from a randomised controlled trial. *International Journal of Obesity,* 41**,** 246-254.

BELL, M. L., FIERO, M., HORTON, N. J. & HSU, C.-H. 2014. Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology,* 14**,** 118.

BENDER, R. & LANGE, S. 2001. Adjusting for multiple testing—when and how? *Journal of clinical epidemiology,* 54**,** 343-349.

BLAKESLEY, R. E., MAZUMDAR, S., DEW, M. A., HOUCK, P. R., TANG, G., REYNOLDS, C. F., 3RD & BUTTERS, M. A. 2009. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology,* 23**,** 255-64.

BRETZ, F., HOTHORN, T. & WESTFALL, P. 2010. *Multiple comparisons using R*, CRC Press.

BRETZ, F., MAURER, W., BRANNATH, W. & POSCH, M. 2009. A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine,* 28**,** 586-604.

BRETZ, F., POSCH, M., GLIMM, E., KLINGLMUELLER, F., MAURER, W. & ROHMEYER, K. 2011. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal,* 53**,** 894-913.

BROOKES, S. T., WHITLEY, E., PETERS, T. J., MULHERAN, P. A., EGGER, M. & DAVEY SMITH, G. 2001. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health technology assessment,* 5**,** 1-56.

BURTON, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2006. The design of simulation studies in medical statistics. *Statistics in medicine,* 25**,** 4279-4292.

BUSZEWICZ, M., GRIFFIN, M., MCMAHON, E. M., BEECHAM, J. & KING, M. 2010. Evaluation of a system of structured, pro-active care for chronic depression in primary care: a randomised controlled trial. *BMC Psychiatry,* 10**,** 61.

BUSZEWICZ, M., GRIFFIN, M., MCMAHON, E. M., WALTERS, K. & KING, M. 2016. Practice nurse-led proactive care for chronic depression in primary care: a randomised controlled trial. *The British Journal of Psychiatry,* 208**,** 374-380.

CANDES, E. 2012. Multiple testng problem lecture 2012. [Accessed 15.07.2013].

CAPIZZI, T. & ZHANG, J. 1996. Testing the hypothesis that matters for multiple primary endpoints. *Drug information journal,* 30**,** 949-956.

CARPENTER, J. R. & KENWARD, M. G. 2007. Missing data in randomised controlled trials: a practical guide. Health Technology Assessment Methodology Programme.

CAUGHEY, D. & CAUGHEY, M. D. 2016. Package 'NPC'. *R Software Package Report*.

CHAN, A.-W., TETZLAFF, J. M., ALTMAN, D. G., LAUPACIS, A., GØTZSCHE, P. C., KRLEŽA-JERIĆ, K., HRÓBJARTSSON, A., MANN, H., DICKERSIN, K. & BERLIN, J. A. 2013. SPIRIT 2013

statement: defining standard protocol items for clinical trials. *Annals of internal medicine,* 158**,** 200-207.

CHOW, S.-C., SHAO, J., WANG, H. & LOKHNYGINA, Y. 2017. *Sample size calculations in clinical research*, Chapman and Hall/CRC.

COLLINS, L. M., SCHAFER, J. L. & KAM, C.-M. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods,* 6**,** 330.

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS 2002. Points to consider on multiplicity issues in clinical trials. *London: The European Agency for the Evaluation of Medicinal Products*.

CONROD, P. J., O'LEARY-BARRETT, M., NEWTON, N., TOPPER, L., CASTELLANOS-RYAN, N., MACKIE, C. & GIRARD, A. 2013. Effectiveness of a selective, personality-targeted prevention program for adolescent alcohol use and misuse: a cluster randomized controlled trial. *JAMA Psychiatry,* 70**,** 334-42.

CORDOBA, G., SCHWARTZ, L., WOLOSHIN, S., BAE, H. & GØTZSCHE, P. C. 2010. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BmJ,* 341**,** c3920.

COX, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological),* 34**,** 187-202.

CROWTHER, M. J., ABRAMS, K. R. & LAMBERT, P. C. 2013. Joint modeling of longitudinal and survival data. *Stata J,* 13**,** 165-184.

CUTTER, G. R., BAIER, M. L., RUDICK, R. A., COOKFAIR, D. L., FISCHER, J. S., PETKAU, J., SYNDULKO, K., WEINSHENKER, B. G., ANTEL, J. P. & CONFAVREUX, C. 1999. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain,* 122**,** 871-882.

DE LOS REYES, A., KUNDEY, S. M. A. & WANG, M. 2011. The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology Review,* 31**,** 829-838.

DMITRIENKO, A. & D'AGOSTINO, R. 2013. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine,* 32**,** 5172-5218.

DMITRIENKO, A., TAMHANE, A. C. & BRETZ, F. 2009. *Multiple testing problems in pharmaceutical statistics*, CRC Press.

DODEL, R., ROMINGER, A., BARTENSTEIN, P., BARKHOF, F., BLENNOW, K., FÖRSTER, S., WINTER, Y., BACH, J.-P., POPP, J. & ALFERINK, J. 2013. Intravenous immunoglobulin for treatment of mild-to-moderate Alzheimer's disease: a phase 2, randomised, double-blind, placebo-controlled, dose-finding trial. *The Lancet Neurology*.

DUNNETT, C. W. 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association,* 50**,** 1096-1121.

DUNSON, D. B. 2000. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 62**,** 355-366.

DWAN, K., KIRKHAM, J. J., WILLIAMSON, P. R. & GAMBLE, C. 2013. Selective reporting of outcomes in randomised controlled trials in systematic reviews of cystic fibrosis. *BMJ open,* 3**,** e002709.

EUROPEAN MEDICAL AGENCY 2017. Guideline on multiplicity issues in clinical trials.

FAUCETT, C. L. & THOMAS, D. C. 1996. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in medicine,* 15**,** 1663-1685.

FERREIRA-GONZÁLEZ, I., PERMANYER-MIRALDA, G., BUSSE, J. W., BRYANT, D. M., MONTORI, V. M., ALONSO-COELLO, P., WALTER, S. D. & GUYATT, G. H. 2007. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of clinical epidemiology,* 60**,** 651-657.

FOOD AND DRUG ADMINISTRATION 2017. Multiple Endpoints in Clinical Trials Guidance for Industry. *FDA Issues Draft Guidance*.

FREEMANTLE, N., CALVERT, M., WOOD, J., EASTAUGH, J. & GRIFFIN, C. 2003. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *Jama,* 289**,** 2554-2559.

FREIDLIN, B., KORN, E. L., GRAY, R. & MARTIN, A. 2008. Multi-arm clinical trials of new agents: some design considerations. *Clinical Cancer Research,* 14**,** 4368-4371.

GE, Y., DUDOIT, S. & SPEED, T. P. 2003. Resampling-based multiple testing for microarray data analysis. *Test,* 12**,** 1-77.

GOLDSTEIN, H. 2011. *Multilevel statistical models*, Wiley. com.

GOLDSTEIN, H., CARPENTER, J., KENWARD, M. G. & LEVIN, K. A. 2009. Multilevel models with multivariate mixed response types. *Statistical Modelling,* 9**,** 173-197.

GOLDSTEIN, H., RASBASH, J., YANG, M., WOODHOUSE, G., PAN, H., NUTTALL, D. & THOMAS, S. 1993. A multilevel analysis of school examination results. *Oxford review of education,* 19**,** 425-433.

GRAHAM, J. W., OLCHOWSKI, A. E. & GILREATH, T. D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science,* 8**,** 206-213.

GRAY, K. M., CARPENTER, M. J., BAKER, N. L., DESANTIS, S. M., KRYWAY, E., HARTWELL, K. J., MCRAE-CLARK, A. L. & BRADY, K. T. 2012. A double-blind randomized controlled trial of N-acetylcysteine in cannabis-dependent adolescents. *American Journal of Psychiatry,* 169**,** 805-812.

GRILLI, L. & RAMPICHINI, C. 2006. A review of random effects modelling using gllamm in Stata. *Department of Statistics, University of Florence*.

HASSIOTIS, A., POPPE, M., STRYDOM, A., VICKERSTAFF, V., HALL, I. S., CRABTREE, J., OMAR, R. Z., KING, M., HUNTER, R. & BISWAS, A. 2018. Clinical outcomes of staff training in positive behaviour support to reduce challenging behaviour in adults with intellectual disability: cluster randomised controlled trial. *The British Journal of Psychiatry,* 212**,** 161-168.

HASSIOTIS, A., ROBOTHAM, D., CANAGASABEY, A., ROMEO, R., LANGRIDGE, D., BLIZARD, R., MURAD, S. & KING, M. 2009. Randomized, Single-Blind, Controlled Trial of a Specialist Behavior Therapy Team for Challenging Behavior in Adults With Intellectual Disabilities. *American Journal of Psychiatry,* 166**,** 1278-1285.

HASSIOTIS, A., STRYDOM, A., CRAWFORD, M., HALL, I., OMAR, R., VICKERSTAFF, V., HUNTER, R., CRABTREE, J., COOPER, V. & BISWAS, A. 2014. Clinical and cost effectiveness of staff training in Positive Behaviour Support (PBS) for treating challenging behaviour in adults with intellectual disability: a cluster randomised controlled trial. *BMC Psychiatry,* 14**,** 219.

HATFIELD, L. A., BOYE, M. E. & CARLIN, B. P. 2011. Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of biopharmaceutical statistics,* 21**,** 971-991.

HENDERSON, R., DIGGLE, P. & DOBSON, A. 2000. Joint modelling of longitudinal measurements and event time data. *Biostatistics,* 1**,** 465-480.

HEO, M. & LEON, A. C. 2008. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics,* 64**,** 1256-1262.

HIDALGO, B. & GOODMAN, M. 2013. Multivariate or multivariable regression? *American journal of public health,* 103**,** 39-40.

HOCHBERG, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* 75**,** 800-802.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics***,** 65-70.

HOMMEL, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika,* 75**,** 383-386.

HONG, L. E., THAKER, G. K., MCMAHON, R. P., SUMMERFELT, A., RACHBEISEL, J., FULLER, R. L., WONODI, I., BUCHANAN, R. W., MYERS, C. & HEISHMAN, S. J. 2011. Effects of moderate-dose treatment with varenicline on neurobiological and cognitive biomarkers in smokers and nonsmokers with schizophrenia or schizoaffective disorder. *Archives of general psychiatry,* 68**,** 1195-1206.

HOPEWELL, S., CLARKE, M., MOHER, D., WAGER, E., MIDDLETON, P., ALTMAN, D. G., SCHULZ, K. F. & GROUP, C. 2008. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS medicine,* 5**,** e20.

HSIEH, F., TSENG, Y. K. & WANG, J. L. 2006. Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics,* 62**,** 1037-1043.

IBRAHIM, J. G., CHU, H. & CHEN, L. M. 2010. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology,* 28**,** 2796.

ICH E9 EXPERT WORKING GROUP 1999. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine,* 18**,** 1903-1942.

JOHNSON, S., RAINS, L. S., MARWAHA, S., STRANG, J., CRAIG, T., WEAVER, T., MCCRONE, P., KING, M., FOWLER, D. & PILLING, S. 2016. A randomised controlled trial of the clinical and cost-effectiveness of a contingency management intervention compared to treatment as usual for reduction of cannabis use and of relapse in early psychosis (CIRCLE): a study protocol for a randomised controlled trial. *Trials,* 17**,** 515.

KAY, S. R., FISZBEIN, A. & OPLER, L. A. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin,* 13**,** 261-276.

KILLASPY, H., MARSTON, L., GREEN, N., HARRISON, I., LEAN, M., COOK, S., MUNDY, T., CRAIG, T., HOLLOWAY, F. & LEAVEY, G. 2015. Clinical effectiveness of a staff training intervention in mental health inpatient rehabilitation units designed to increase patients' engagement in activities (the Rehabilitation Effectiveness for Activities for Life [REAL] study): single-blind, cluster-randomised controlled trial. *The Lancet Psychiatry,* 2**,** 38-48.

KIM, S., ZENG, D., CHAMBLESS, L. & LI, Y. 2012. Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in biosciences,* 4**,** 262-281.

KING, M., DAVIDSON, O., TAYLOR, F., HAINES, A., SHARP, D. & TURNER, R. 2002. Effectiveness of teaching general practitioners skills in brief cognitive behaviour therapy to treat patients with depression: randomised controlled trial. *Bmj,* 324**,** 947.

KOHL, M. & KOLAMPALLY, S. 2017. mpe: Multiple Primary Endpoints.

KRÓL, A., MAUGUEN, A., MAZROUI, Y., LAURENT, A., MICHIELS, S. & RONDEAU, V. 2017. Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *arXiv preprint arXiv:1701.03675*.

LAFAYE DE MICHEAUX, P., LIQUET, B., MARQUE, S. & RIOU, J. 2014. Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. *Journal of biopharmaceutical statistics,* 24**,** 378-397.

LAUNER, L. J., MILLER, M. E., WILLIAMSON, J. D., LAZAR, R. M., GERSTEIN, H. C., MURRAY, A. M., SULLIVAN, M., HOROWITZ, K. R., DING, J. & MARCOVINA, S. 2011. Effects of intensive glucose lowering on brain structure and function in people with type 2 diabetes (ACCORD MIND): a randomised open-label substudy. *The Lancet Neurology,* 10**,** 969-977.

LAWRENCE GOULD, A., BOYE, M. E., CROWTHER, M. J., IBRAHIM, J. G., QUARTEY, G., MICALLEF, S. & BOIS, F. Y. 2015. Joint modeling of survival and longitudinal non-

survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in medicine,* 34**,** 2181-2195.

LECKIE, G. & CHARLTON, C. 2013. Runmlwin-a program to Run the MLwiN multilevel modelling software from within stata. *Journal of Statistical Software,* 52**,** 1-40.

LEISCH, F., WEINGESSEL, A. & HORNIK, K. 1998. On the generation of correlated artificial binary data.

LI, D. & DYE, T. D. 2013. Power and stability properties of resampling-based multiple testing procedures with applications to gene oncology studies. *Computational and mathematical methods in medicine,* 2013.

LI, J. D. & MEHROTRA, D. V. 2008. An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in medicine,* 27**,** 5377-5391.

LITTLE, R. J., D'AGOSTINO, R., COHEN, M. L., DICKERSIN, K., EMERSON, S. S., FARRAR, J. T., FRANGAKIS, C., HOGAN, J. W., MOLENBERGHS, G. & MURPHY, S. A. 2012. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine,* 367**,** 1355-1360.

LITTLE, R. J. & RUBIN, D. B. 2014. *Statistical analysis with missing data*, John Wiley & Sons.

LIU, L., WOLFE, R. A. & HUANG, X. 2004. Shared frailty models for recurrent events and a terminal event. *Biometrics,* 60**,** 747-756.

LOVERA, J. F., KIM, E., HERIZA, E., FITZPATRICK, M., HUNZIKER, J., TURNER, A. P., ADAMS, J., STOVER, T., SANGEORZAN, A. & SLOAN, A. 2012. Ginkgo biloba does not improve cognitive function in MS A randomized placebo-controlled trial. *Neurology,* 79**,** 1278-1284.

MAAS, J., VERHEIJ, R. A., SPREEUWENBERG, P. & GROENEWEGEN, P. P. 2008. Physical activity as a possible mechanism behind the relationship between green space and health: a multilevel analysis. *BMC public health,* 8**,** 206.

MACHIN, D., CAMPBELL, M. J., TAN, S. B. & TAN, S. H. 2018. *Sample Sizes for Clinical, Laboratory and Epidemiology Studies*, Wiley Online Library.

MAYO, N. E. & SCOTT, S. 2011. Evaluating a complex intervention with a single outcome may not be a good idea: an example from a randomised trial of stroke case management. *Age and ageing,* 40**,** 718-724.

MAZROUI, Y., MATHOULIN-PELISSIER, S., SOUBEYRAN, P. & RONDEAU, V. 2012. General joint frailty model for recurrent event data with a dependent terminal event: application to follicular lymphoma data. *Statistics in medicine,* 31**,** 1162-1176.

MCCULLOCH, C. 2008a. Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research,* 17**,** 53-73.

MCCULLOCH, C. 2008b. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res,* 17**,** 53-73.

MOHAN, J., TWIGG, L. & TAYLOR, J. 2011. Mind The Double Gap Using Multivariate Multilevel Modelling to Investigate Public Perceptions of Crime Trends. *British Journal of Criminology,* 51**,** 1035-1053.

MONTORI, V. M., PERMANYER-MIRALDA, G., FERREIRA-GONZÁLEZ, I., BUSSE, J. W., PACHECO-HUERGO, V., BRYANT, D., ALONSO, J., AKL, E. A., DOMINGO-SALVANY, A. & MILLS, E. 2005. Validity of composite end points in clinical trials. *BMJ: British Medical Journal,* 330**,** 594.

MOYÉ, L. A. 2000. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in medicine,* 19**,** 767-779.

MOYÉ, L. A. 2003. *Multiple analyses in clinical trials: fundamentals for investigators*, Springer Science & Business Media.

NIERENBERG, A. A., FRIEDMAN, E. S., BOWDEN, C. L., SYLVIA, L. G., THASE, M. E., KETTER, T., OSTACHER, M. J., LEON, A. C., REILLY-HARRINGTON, N. & IOSIFESCU, D. V. 2013. Lithium Treatment Moderate-Dose Use Study (LiTMUS) for bipolar disorder: a

randomized comparative effectiveness trial of optimized personalized treatment with and without lithium. *American Journal of Psychiatry,* 170**,** 102-110.

NIXON, R. M. & THOMPSON, S. G. 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health economics,* 14**,** 1217-1229.

NOBILE-ORAZIO, E., COCITO, D., JANN, S., UNCINI, A., BEGHI, E., MESSINA, P., ANTONINI, G., FAZIO, R., GALLIA, F. & SCHENONE, A. 2012. Intravenous immunoglobulin versus intravenous methylprednisolone for chronic inflammatory demyelinating polyradiculoneuropathy: a randomised controlled trial. *The Lancet Neurology,* 11**,** 493-502.

ODEKERKEN, V. J., VAN LAAR, T., STAAL, M. J., MOSCH, A., HOFFMANN, C. F., NIJSSEN, P. C., BEUTE, G. N., VAN VUGT, J. P., LENDERS, M. W. & CONTARINO, M. F. 2012. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. *The Lancet Neurology*.

OFFEN, W., CHUANG-STEIN, C., DMITRIENKO, A., LITTMAN, G., MACA, J., MEYERSON, L., MUIRHEAD, R., STRYSZAK, P., BADDY, A. & CHEN, K. 2007. Multiple co-primary endpoints: medical and statistical solutions: a report from the multiple endpoints expert team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal,* 41**,** 31-46.

OSBORN, D. P., HARDOON, S., OMAR, R. Z., HOLT, R. I., KING, M., LARSEN, J., MARSTON, L., MORRIS, R. W., NAZARETH, I. & WALTERS, K. 2015. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. *JAMA psychiatry,* 72**,** 143-151.

PAUX, G. & DMITRIENKO, A. 2018. Package 'Mediana': Clinical Trial Simulations. 1.0.7 ed.

PHILIPSON, P., DIGGLE, P., SOUSA, I., KOLAMUNNAGE-DONA, R., WILLIAMSON, P. & HENDERSON, R. 2012. joineR: Joint modelling of repeated measurements and time-to-event data.

PHILLIPS, A. & HAUDIQUET, V. 2003. ICH E9 guideline 'Statistical principles for clinical trials': a case study. *Statistics in Medicine,* 22**,** 1-11.

PITUCH, K. A., WHITTAKER, T. A. & CHANG, W. 2016. Multivariate Models for Normal and Binary Responses in Intervention Studies. *American Journal of Evaluation,* 37**,** 270-286.

POCOCK, S. 1997. Clinical trials with multiple outcomes - a statistical perespsective on their design, analysis and interpration. *Controlled Clinical Trials,* 18**,** 530-545.

POCOCK, S., GELLER, N. & TSIATIS, A. 1987. The analysis of multiple endpoints in clinical trials. *Biometrics,* 43**,** 487-498.

POGUE, J., DEVEREAUX, P., THABANE, L. & YUSUF, S. 2012. Designing and analyzing clinical trials with composite outcomes: consideration of possible treatment differences between the individual outcomes. *PloS one,* 7**,** e34785.

RABASH, J., CHARLTON, C., BROWNE, W., HEALY, M. & CAMERON, B. 2009. Mlwin version 2.1. Centre for Multilevel Modelling. *University of Bristol*.

RABE-HESKETH, S. & SKRONDAL, A. 2008. *Multilevel and longitudinal modeling using Stata*, STATA press.

RABE-HESKETH, S., SKRONDAL, A. & PICKLES, A. 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics,* 128**,** 301-323.

RASBASH, J., STEELE, F., BROWNE, W. J. & GOLDSTEIN, H. 2012. A user's guide to MLwiN Version 2.26. Centre for Multilevel Modelling, University of Bristol.

REITMEIR, P. & WASSMER, G. 1999. Resampling-based methods for the analysis of multiple endpoints in clinical trials. *Statistics in Medicine,* 18**,** 3453-3462.

RIZOPOULOS, D. 2010. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online),* 35**,** 1-33.

RÖHRIG, B., DU PREL, J.-B., WACHTLIN, D., KWIECIEN, R. & BLETTNER, M. 2010. Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International,* 107**,** 552.

RONDEAU, V., MAZROUI, Y. & GONZALEZ, J. R. 2012. frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw,* 47**,** 1-28.

ROTHWELL, J. C., JULIOUS, S. A. & COOPER, C. L. 2018. A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal. *Trials,* 19**,** 544.

ROY, J., LIN, X. & RYAN, L. M. 2003. Scaled marginal models for multiple continuous outcomes. *Biostatistics,* 4**,** 371-383.

RUBIN, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American statistical Association,* 91**,** 473-489.

RUBIN, D. B. 2004. *Multiple imputation for nonresponse in surveys*, John Wiley & Sons.

SAMMEL, M. D., RYAN, L. M. & LEGLER, J. M. 1997. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 59**,** 667-678.

SANKOH, A. J., HUQUE, M. F. & DUBEY, S. D. 1997. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in medicine,* 16**,** 2529-2542.

SCHAFER, J. L. & GRAHAM, J. W. 2002. Missing data: our view of the state of the art. *Psychological methods,* 7**,** 147.

SCHERER, R. 2016. samplesize: Sample Size Calculation for Various t-Tests and Wilcoxon-Test.

SCHULZ, K. F., ALTMAN, D. G. & MOHER, D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine,* 8**,** 18.

SEDDON, J. L., BIRCHWOOD, M., COPELLO, A., EVERARD, L., JONES, P. B., FOWLER, D., AMOS, T., FREEMANTLE, N., SHARMA, V. & MARSHALL, M. 2015. Cannabis use is associated with increased psychotic symptoms and poorer psychosocial functioning in first-episode psychosis: a report from the UK national EDEN study. *Schizophrenia bulletin,* 42**,** 619-625.

SENDYK, D. I., ROVAI, E. S., SOUZA, N. V., DEBONI, M. C. Z. & PANNUTI, C. M. 2019. Selective outcome reporting in randomized clinical trials of dental implants. *Journal of clinical periodontology,* 46**,** 758-765.

SENN, S. & BRETZ, F. 2007. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics,* 6**,** 161-170.

SHAFFER, J. P. 1995. Multiple hypothesis testing. *Annual review of psychology,* 46**,** 561-584.

ŠIDÁK, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association,* 62**,** 626-633.

SIMES, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika,* 73**,** 751-754.

SINHARAY, S., STERN, H. S. & RUSSELL, D. 2001. The use of multiple imputation for the analysis of missing data. *Psychological methods,* 6**,** 317.

SKRONDAL, A. & RABE-HESKETH, S. 2004. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, Crc Press.

SNIJDERS, T. & BOSKER, R. J. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second edition. ,* London Sage Publishers.

SOZU, T., SUGIMOTO, T., HAMASAKI, T. & EVANS, S. R. 2015. *Sample size determination in clinical trials with multiple endpoints*, Springer.

SPIEGELHALTER, D. J., ABRAMS, K. R. & MYLES, J. P. 2004. *Bayesian approaches to clinical trials and health-care evaluation*, John Wiley & Sons.

STATACORP, L. 2015. Stata Statistical Software: Release 14. *Texas USA: StataCorp LP*.

STATACORP STATACORP, L. College Station, TX: 2011. *Stata statistical software: release,* 12.

STERNE, J. A., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj,* 338**,** b2393.

SUGIMOTO, T., SOZU, T. & HAMASAKI, T. 2012. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. *Pharmaceutical Statistics,* 11**,** 118-128.

SULLIVAN, T. R., WHITE, I. R., SALTER, A. B., RYAN, P. & LEE, K. J. 2018. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical methods in medical research,* 27**,** 2610-2626.

TARIOT, P. N., SCHNEIDER, L. S., CUMMINGS, J., THOMAS, R. G., RAMAN, R., JAKIMOVICH, L. J., LOY, R., BARTOCCI, B., FLEISHER, A. & ISMAIL, M. S. 2011. Chronic divalproex sodium to attenuate agitation and clinical progression of Alzheimer disease. *Archives of General Psychiatry,* 68**,** 853.

TEIXEIRA-PINTO, A. & HAREZLAK, J. 2013. Factorization and latent variable models for joint analysis of binary and continuous outcomes. *Analysis of Mixed Data: Methods & Applications, CRC Press, Taylor & Francis Group, Boca Raton, FL,* 81**,** 91.

TEIXEIRA-PINTO, A., SIDDIQUE, J., GIBBONS, R. & NORMAND, S.-L. 2009. Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric annals,* 39**,** 729.

TEIXEIRA-PINTO, A. & NORMAND, S. L. T. 2008. Statistical methodology for classifying units on the basis of multiple-related measures. *Statistics in medicine,* 27**,** 1329-1350.

TEIXEIRA-PINTO, A. & NORMAND, S. L. T. 2009. Correlated bivariate continuous and binary outcomes: issues and applications. *Statistics in medicine,* 28**,** 1753-1773.

THOMPSON, S. G. & NIXON, R. M. 2005. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Medical Decision Making,* 25**,** 416-423.

TSELONI, A. & ZARAFONITOU, C. 2008. Fear of crime and victimization: A multivariate multilevel analysis of competing measurements. *European Journal of Criminology,* 5**,** 387-409.

TUKEY, J., CIMINERA, J. & HEYSE, J. 1985. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics***,** 295-301.

TYLER, K. M., NORMAND, S. L. & HORTON, N. J. 2011. The use and abuse of multiple outcomes in randomized controlled depression trials. *Contemp Clin Trials,* 32**,** 299-304.

VICKERSTAFF, V., AMBLER, G., KING, M., NAZARETH, I. & OMAR, R. Z. 2015. Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. *Contemporary clinical trials,* 45**,** 8-12.

VICKERSTAFF, V., OMAR, R. Z. & AMBLER, G. 2019. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC medical research methodology,* 19**,** 129.

VITIELLO, B., ELLIOTT, G. R., SWANSON, J. M., ARNOLD, L. E., HECHTMAN, L., ABIKOFF, H., MOLINA, B. S., WELLS, K., WIGAL, T. & JENSEN, P. S. 2014. Blood pressure and heart rate over 10 years in the multimodal treatment study of children with ADHD. *American Journal of Psychiatry*.

WANG, P., SHEN, W. & BOYE, M. E. 2012. Joint modeling of longitudinal outcomes and survival using latent growth modeling approach in a mesothelioma trial. *Health Services and Outcomes Research Methodology,* 12**,** 182-199.

WANG, Y. & TAYLOR, J. M. G. 2001. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association,* 96**,** 895-905.

WARNER, R. M. 2008. *Applied statistics: From bivariate through multivariate techniques*, Sage.

WEAVER, F. M., FOLLETT, K. A., STERN, M., LUO, P., HARRIS, C. L., HUR, K., MARKS, W. J., ROTHLIND, J., SAGHER, O. & MOY, C. 2012. Randomized trial of deep brain stimulation for Parkinson disease Thirty-six-month outcomes. *Neurology,* 79**,** 55-65.

WEISS, R. D., POTTER, J. S., FIELLIN, D. A., BYRNE, M., CONNERY, H. S., DICKINSON, W., GARDIN, J., GRIFFIN, M. L., GOUREVITCH, M. N. & HALLER, D. L. 2011. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence: a 2-phase randomized controlled trial. *Archives of General Psychiatry***,** archgenpsychiatry. 2011.121 v1.

WESTFALL, P. H., TOBIAS, R. D. & WOLFINGER, R. D. 2011. *Multiple comparisons and multiple tests using SAS*, SAS Institute.

WESTFALL, P. H. & YOUNG, S. S. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, John Wiley & Sons.

WHITE, I. R. & CARLIN, J. B. 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine,* 29**,** 2920-2931.

WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine,* 30**,** 377-399.

WHO, W. H. O. 2012. International Standards for Clinical Trial Registries. . http://www.who.int/iris/bitstream/10665/76705/1/9789241504294_eng.pdf?ua=1 .

WITTCHEN, H.-U., JACOBI, F., REHM, J., GUSTAVSSON, A., SVENSSON, M., JÖNSSON, B., OLESEN, J., ALLGULANDER, C., ALONSO, J. & FARAVELLI, C. 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology,* 21**,** 655-679.

WOOD, A. M., WHITE, I. R. & THOMPSON, S. G. 2004. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical trials,* 1**,** 368-376.

WRIGHT, S. P. 1992. Adjusted p-values for simultaneous inference. *Biometrics***,** 1005-1013.

YANG, M., GOLDSTEIN, H., BROWNE, W. & WOODHOUSE, G. 2002. Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 165**,** 137-153.

YOON, F. B., FITZMAURICE, G. M., LIPSITZ, S. R., HORTON, N. J., LAIRD, N. M. & NORMAND, S. L. 2011. Alternative methods for testing treatment effects on the basis of multiple outcomes: simulation and case study. *Statistics in Medicine,* 30**,** 1917-32.

YUEN, H. P. & MACKINNON, A. 2016. Performance of joint modelling of time-to-event data with time-dependent predictors: an assessment based on transition to psychosis data. *PeerJ,* 4**,** e2582.

ZHANG, D., CHEN, M.-H., IBRAHIM, J. G., BOYE, M. E. & SHEN, W. 2016a. JMFit: a SAS macro for joint models of longitudinal and survival data. *Journal of statistical software,* 71.

ZHANG, Z., PARKER, R., CHARLTON, C. M., LECKIE, G. & BROWNE, W. J. 2016b. R2MLwiN: A package to run MLwiN from within R. *Journal of Statistical Software,* 72.

List of appendices

Appendix 1

Published manuscript for the review of neurology and psychiatry randomised controlled trials

Contemporary Clinical Trials 45 (2015) 8–12

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review

Victoria Vickerstaff [a,b,c,*], Gareth Ambler [b], Michael King [a], Irwin Nazareth [c], Rumana Z. Omar [b]

[a] Division of Psychiatry, University College London, 6th Floor, Maple House, 149 Tottenham Court Road, London W1T 7NF, UK
[b] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
[c] The Research Department of Primary Care and Population Health, University College London, Rowland Hill Street, London NW3 2PF, UK

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form 18 July 2015
Accepted 20 July 2015
Available online 26 July 2015

Keywords:
Multiplicity
Multiple outcomes
Clinical trials
Neurology
Psychiatry

ABSTRACT

Objectives: To review how multiple primary outcomes are currently considered in the analysis of randomised controlled trials. We briefly describe the methods available to safeguard the inferences and to raise awareness of the potential problems caused by multiple outcomes.
Methods/design: We reviewed randomised controlled trials (RCTs) in neurology and psychiatry disease areas, as these frequently analyse multiple outcomes. We reviewed all published RCTs from July 2011 to June 2014 inclusive in the following high impact journals: The New England Journal of Medicine, The Lancet, The American Journal of Psychiatry, JAMA Psychiatry, The Lancet Neurology and Neurology. We examined the information presented in the abstract and the methods used for sample size calculation and statistical analysis. We recorded the number of primary outcomes, the methods used to account for multiple primary outcomes, the number of outcomes discussed in the abstract and the number of outcomes used in the sample size calculation.
Results: Of the 209 RCTs that we identified, 60 (29%) analysed multiple primary outcomes. Of these, 45 (75%) did not adjust for multiplicity in their analyses. Had multiplicity been addressed, some of the trial conclusions would have changed. Of the 15 (25%) trials which accounted for multiplicity, Bonferroni's correction was the most commonly used method.
Conclusions: Our review shows that trials with multiple primary outcomes are common. However, appropriate steps are not usually taken in most of the analyses to safeguard the inferences against multiplicity. Authors should state their chosen primary outcomes clearly and justify their methods of analysis.

http://dx.doi.org/10.1016/j.cct.2015.07.016
1551-7144/

Appendix 2

Papers included in review of published neurology and psychiatry randomised controlled trials

Below are the references for the 209 trials in the review of recently published randomised controlled trials (Chapter 3)

1. ADELSON, P. D., WISNIEWSKI, S. R., BECA, J., BROWN, S. D., BELL, M., MUIZELAAR, J. P., OKADA, P., BEERS, S. R., BALASUBRAMANI, G. K. & HIRTZ, D. 2013. Comparison of hypothermia and normothermia after severe traumatic brain injury in children (Cool Kids): a phase 3, randomised controlled trial. The Lancet Neurology.
2. AL-CHALABI, A., ALLEN, C., COUNSELL, C., FARRIN, A., DICKIE, B., KELLY, J., LEIGH, P., MURPHY, C., PAYAN, C. & REYNOLDS, G. 2013. Lithium in patients with amyotrophic lateral sclerosis (LiCALS): A phase 3 multicentre, randomised, double-blind, placebo-controlled trial. The Lancet Neurology, 12, 339-345.
3. ALEGRÍA, M., CARSON, N., FLORES, M., LI, X., SHI, P., LESSIOS, A. S., POLO, A., ALLEN, M., FIERRO, M. & INTERIAN, A. 2014. Activation, self-management, engagement, and retention in behavioural health care: a randomized clinical trial of the DECIDE intervention. JAMA psychiatry.
4. ALLEN, R. P., CHEN, C., GARCIA-BORREGUERO, D., POLO, O., DUBRAVA, S., MICELI, J., KNAPP, L. & WINKELMAN, J. W. 2014. Comparison of pregabalin with pramipexole for restless legs syndrome. New England Journal of Medicine, 370, 621-631.
5. ALLSOP, D. J., COPELAND, J., LINTZERIS, N., DUNLOP, A. J., MONTEBELLO, M., SADLER, C., RIVAS, G. R., HOLLAND, R. M., MUHLEISEN, P. & NORBERG, M. M. 2014. Nabiximols as an Agonist Replacement Therapy During Cannabis Withdrawal: A Randomized Clinical Trial. JAMA psychiatry, 71, 281-291.
6. ALTINBAS, A., VAN ZANDVOORT, M., VAN DEN BERG, E., JONGEN, L., ALGRA, A., MOLL, F., NEDERKOORN, P., MALI, W., BONATI, L. & BROWN, M. 2011. Cognition after carotid endarterectomy or stenting A randomized comparison. Neurology, 77, 1084-1090.
7. ANDERSON, C. S., HEELEY, E., HUANG, Y., WANG, J., STAPF, C., DELCOURT, C., LINDLEY, R., ROBINSON, T., LAVADOS, P. & NEAL, B. 2013. Rapid blood-pressure lowering in patients with acute intracerebral haemorrhage. New England Journal of Medicine, 368, 2355-2365.
8. ANTON, R. F., MYRICK, H., WRIGHT, T. M., LATHAM, P. K., BAROS, A. M., WAID, L. R. & RANDALL, P. K. 2011. Gabapentin combined with naltrexone for the treatment of alcohol dependence. American Journal of Psychiatry, 168, 709-717.
9. BAMELIS, L. L., EVERS, S. M., SPINHOVEN, P. & ARNTZ, A. 2014. Results of a multicentre randomized controlled trial of the clinical effectiveness of Schema Therapy for personality disorders. American Journal of Psychiatry, 171, 305-322.
10. BANERJEE, S., HELLIER, J., DEWEY, M., ROMEO, R., BALLARD, C., BALDWIN, R., BENTHAM, P., FOX, C., HOLMES, C. & KATONA, C. 2011. Sertraline or mirtazapine for depression in dementia (HTA-SADD): a randomised, multicentre, double-blind, placebo-controlled trial. The Lancet, 378, 403-411.
11. BARLOW, A., MULLANY, B., NEAULT, N., COMPTON, S., CARTER, A., HASTINGS, R., BILLY, T., COHO-MESCAL, V., LORENZO, S. & WALKUP, J. T. 2013. Effect of a paraprofessional home-

visiting intervention on American Indian teen mothers' and infants' behavioural risks: A randomized controlled trial. American Journal of Psychiatry, 170, 83-93.

12. BASS, J. K., ANNAN, J., MCIVOR MURRAY, S., KAYSEN, D., GRIFFITHS, S., CETINOGLU, T., WACHTER, K., MURRAY, L. K. & BOLTON, P. A. 2013. Controlled trial of psychotherapy for Congolese survivors of sexual violence. New England Journal of Medicine, 368, 2182-2191.
13. BAULAC, M., BRODIE, M. J., PATTEN, A., SEGIETH, J. & GIORGI, L. 2012. Efficacy and tolerability of zonisamide versus controlled-release carbamazepine for newly diagnosed partial epilepsy: a phase 3, randomised, double-blind, non-inferiority trial. The Lancet Neurology, 11, 579-588.
14. BEARDSLEE, W. R., BRENT, D. A., WEERSING, V. R., CLARKE, G. N., PORTA, G., HOLLON, S. D., GLADSTONE, T. R., GALLOP, R., LYNCH, F. L. & IYENGAR, S. 2013. Prevention of depression in at-risk adolescents: Longer-term effects. JAMA psychiatry, 70, 1161-1170.
15. BENAVENTE, O., COFFEY, C., CONWIT, R., HART, R., MCCLURE, L., PEARCE, L., PERGOLA, P., SZYCHOWSKI, J. & GROUP, S. S. 2013. Blood-pressure targets in patients with recent lacunar stroke: the SPS3 randomised trial. Lancet, 382, 507.
16. BENAVENTE, O. R., HART, R. G., MCCLURE, L. A., SZYCHOWSKI, J. M., COFFEY, C. S., PEARCE, L. A. & INVESTIGATORS, S. 2012. Effects of clopidogrel added to aspirin in patients with recent lacunar stroke. N Engl J Med, 367, 817-25.
17. BOWEN, S., WITKIEWITZ, K., CLIFASEFI, S. L., GROW, J., CHAWLA, N., HSU, S. H., CARROLL, H. A., HARROP, E., COLLINS, S. E. & LUSTYK, M. K. 2014. Relative Efficacy of Mindfulness-Based Relapse Prevention, Standard Relapse Prevention, and Treatment as Usual for Substance Use Disorders: A Randomized Clinical Trial. JAMA psychiatry.
18. BOWIE, C. R., MCGURK, S. R., MAUSBACH, B., PATTERSON, T. L. & HARVEY, P. D. 2012. Combined cognitive remediation and functional skills training for schizophrenia: effects on cognition, functional competence, and real-world behavior. American Journal of Psychiatry, 169, 710-718.
19. BOXER, A. L., KNOPMAN, D. S., KAUFER, D. I., GROSSMAN, M., ONYIKE, C., GRAF-RADFORD, N., MENDEZ, M., KERWIN, D., LERNER, A. & WU, C.-K. 2013. Memantine in patients with frontotemporal lobar degeneration: a multicentre, randomised, double-blind, placebo-controlled trial. The Lancet Neurology, 12, 149-156.
20. BOYER, L., DOUSSET, A., ROUSSEL, P., DOSSETTO, N., CAMMILLERI, S., PIANO, V., KHALFA, S., MUNDLER, O., DONNET, A. & GUEDJ, E. 2014. rTMS in fibromyalgia A randomized trial evaluating QoL and its brain metabolic substrate. Neurology, 82, 1231-1238.
21. BRODERICK, J. P., PALESCH, Y. Y., DEMCHUK, A. M., YEATTS, S. D., KHATRI, P., HILL, M. D., JAUCH, E. C., JOVIN, T. G., YAN, B. & SILVER, F. L. 2013. Endovascular therapy after intravenous t-PA versus t-PA alone for stroke. New England Journal of Medicine, 368, 893-903.
22. BURNS, T., RUGKÅSA, J., MOLODYNSKI, A., DAWSON, J., YEELES, K., VAZQUEZ-MONTES, M., VOYSEY, M., SINCLAIR, J. & PRIEBE, S. 2013. Community treatment orders for patients with psychosis (OCTET): a randomised controlled trial. The Lancet, 381, 1627-1633.
23. CALABRESI, P. A., RADUE, E.-W., GOODIN, D., JEFFERY, D., RAMMOHAN, K. W., REDER, A. T., VOLLMER, T., AGIUS, M. A., KAPPOS, L. & STITES, T. 2014. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. The Lancet Neurology, 13, 545-556.
24. CAMPBELL, A. N., NUNES, E. V., MATTHEWS, A. G., STITZER, M., MIELE, G. M., POLSKY, D., TURRIGIANO, E., WALTERS, S., MCCLURE, E. A., KYLE, T. L., WAHLE, A., VAN VELDHUISEN, P., GOLDMAN, B., BABCOCK, D., STABILE, P. Q., WINHUSEN, T. & GHITZA, U. E. 2014. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. Am J Psychiatry, 171, 683-90.
25. CARDENAS, D. D., NIESHOFF, E. C., SUDA, K., GOTO, S.-I., SANIN, L., KANEKO, T., SPORN, J., PARSONS, B., SOULSBY, M. & YANG, R. 2013. A randomized trial of pregabalin in patients with neuropathic pain due to spinal cord injury. Neurology, 80, 533-539.

26. CARROLL, J. D., SAVER, J. L., THALER, D. E., SMALLING, R. W., BERRY, S., MACDONALD, L. A., MARKS, D. S. & TIRSCHWELL, D. L. 2013. Closure of patent foramen ovale versus medical therapy after cryptogenic stroke. New England Journal of Medicine, 368, 1092-1100.
27. CARROLL, K. M., KILUK, B. D., NICH, C., GORDON, M. A., PORTNOY, G. A., MARINO, D. R. & BALL, S. A. 2014. Computer-assisted delivery of cognitive-behavioral therapy: efficacy and durability of CBT4CBT among cocaine-dependent individuals maintained on methadone. American Journal of Psychiatry, 171, 436-444.
28. CHANCELLOR, M. B., PATEL, V., LENG, W. W., SHENOT, P. J., LAM, W., GLOBE, D. R., LOEB, A. L. & CHAPPLE, C. R. 2013. OnabotulinumtoxinA improves quality of life in patients with neurogenic detrusor overactivity. Neurology, 81, 841-848.
29. CHATTERJEE, S., NAIK, S., JOHN, S., DABHOLKAR, H., BALAJI, M., KOSCHORKE, M., VARGHESE, M., THARA, R., WEISS, H. A. & WILLIAMS, P. 2014. Effectiveness of a community-based intervention for people with schizophrenia and their caregivers in India (COPSI): a randomised controlled trial. The Lancet, 383, 1385-1394.
30. CHESNUT, R. M., TEMKIN, N., CARNEY, N., DIKMEN, S., RONDINA, C., VIDETTA, W., PETRONI, G., LUJAN, S., PRIDGEON, J. & BARBER, J. 2012. A trial of intracranial-pressure monitoring in traumatic brain injury. New England Journal of Medicine, 367, 2471-2481.
31. CHIARAVALLOTI, N. D., MOORE, N. B., NIKELSHPUR, O. M. & DELUCA, J. 2013. An RCT to treat learning impairment in multiple sclerosis The MEMREHAB trial. Neurology, 81, 2066-2072.
32. CICCONE, A., VALVASSORI, L., NICHELATTI, M., SGOIFO, A., PONZIO, M., STERZI, R. & BOCCARDI, E. 2013. Endovascular treatment for acute ischemic stroke. New England Journal of Medicine, 368, 904-913.
33. CINCIRIPINI, P. M., ROBINSON, J. D., KARAM-HAGE, M., MINNIX, J. A., LAM, C., VERSACE, F., BROWN, V. L., ENGELMANN, J. M. & WETTER, D. W. 2013. Effects of varenicline and bupropion sustained-release use plus intensive smoking cessation counseling on prolonged abstinence from smoking and on depression, negative affect, and other symptoms of nicotine withdrawal. JAMA Psychiatry, 70, 522-33.
34. COELHO, T., MAIA, L. F., DA SILVA, A. M., CRUZ, M. W., PLANTÉ-BORDENEUVE, V., LOZERON, P., SUHR, O. B., CAMPISTOL, J. M., CONCEIÇÃO, I. M. & SCHMIDT, H. H.-J. 2012. Tafamidis for transthyretin familial amyloid polyneuropathy A randomized, controlled trial. Neurology, 79, 785-792.
35. COHEN, J. A., COLES, A. J., ARNOLD, D. L., CONFAVREUX, C., FOX, E. J., HARTUNG, H.-P., HAVRDOVA, E., SELMAJ, K. W., WEINER, H. L. & FISHER, E. 2012. Alemtuzumab versus interferon beta 1a as first-line treatment for patients with relapsing-remitting multiple sclerosis: a randomised controlled phase 3 trial. The Lancet, 380, 1819-1828.
36. COLES, A. J., TWYMAN, C. L., ARNOLD, D. L., COHEN, J. A., CONFAVREUX, C., FOX, E. J., HARTUNG, H.-P., HAVRDOVA, E., SELMAJ, K. W. & WEINER, H. L. 2012. Alemtuzumab for patients with relapsing multiple sclerosis after disease-modifying therapy: a randomised controlled phase 3 trial. The Lancet, 380, 1829-1839.
37. COLFAX, G. N., SANTOS, G.-M., DAS, M., SANTOS, D. M., MATHESON, T., GASPER, J., SHOPTAW, S. & VITTINGHOFF, E. 2011. Mirtazapine to reduce methamphetamine use: a randomized controlled trial. Archives of general psychiatry, 68, 1168.
38. COMI, G., DE STEFANO, N., FREEDMAN, M. S., BARKHOF, F., POLMAN, C. H., UITDEHAAG, B. M., CASSET-SEMANAZ, F., HENNESSY, B., MORAGA, M. S. & ROCAK, S. 2012a. Comparison of two dosing frequencies of subcutaneous interferon beta-1a in patients with a first clinical demyelinating event suggestive of multiple sclerosis (REFLEX): a phase 3 randomised controlled trial. The Lancet Neurology, 11, 33-41.
39. COMI, G., JEFFERY, D., KAPPOS, L., MONTALBAN, X., BOYKO, A., ROCCA, M. A. & FILIPPI, M. 2012b. Placebo-controlled trial of oral laquinimod for multiple sclerosis. New England Journal of Medicine, 366, 1000-1009.
40. CONFAVREUX, C., O'CONNOR, P., COMI, G., FREEDMAN, M. S., MILLER, A. E., OLSSON, T. P., WOLINSKY, J. S., BAGULHO, T., DELHAY, J.-L. & DUKOVIC, D. 2014. Oral teriflunomide for

patients with relapsing multiple sclerosis (TOWER): a randomised, double-blind, placebo-controlled, phase 3 trial. The Lancet Neurology, 13, 247-256.

41. CONROD, P. J., O'LEARY-BARRETT, M., NEWTON, N., TOPPER, L., CASTELLANOS-RYAN, N., MACKIE, C. & GIRARD, A. 2013. Effectiveness of a selective, personality-targeted prevention program for adolescent alcohol use and misuse: a cluster randomized controlled trial. JAMA psychiatry, 70, 334-342.
42. CREEMERS, H., VELDINK, J. H., GRUPSTRA, H., NOLLET, F., BEELEN, A. & VAN DEN BERG, L. H. 2014. Cluster RCT of case management on patients' quality of life and caregiver strain in ALS. Neurology, 82, 23-31.
43. CUDKOWICZ, M. E., VAN DEN BERG, L. H., SHEFNER, J. M., MITSUMOTO, H., MORA, J. S., LUDOLPH, A., HARDIMAN, O., BOZIK, M. E., INGERSOLL, E. W. & ARCHIBALD, D. 2013. Dexpramipexole versus placebo for patients with amyotrophic lateral sclerosis (EMPOWER): a randomised, double-blind, phase 3 trial. The Lancet Neurology, 12, 1059-1067.
44. CUMMINGS, J., ISAACSON, S., MILLS, R., WILLIAMS, H., CHI-BURRIS, K., CORBETT, A., DHALL, R. & BALLARD, C. 2014. Pimavanserin for patients with Parkinson's disease psychosis: a randomised, placebo-controlled phase 3 trial. The Lancet, 383, 533-540.
45. DAUVILLIERS, Y., BASSETTI, C., LAMMERS, G. J., ARNULF, I., MAYER, G., RODENBECK, A., LEHERT, P., DING, C.-L., LECOMTE, J.-M. & SCHWARTZ, J.-C. 2013. Pitolisant versus placebo or modafinil in patients with narcolepsy: a double-blind, randomised trial. The Lancet Neurology, 12, 1068-1075.
46. DÁVALOS, A., ALVAREZ-SABÍN, J., CASTILLO, J., DÍEZ-TEJEDOR, E., FERRO, J., MARTÍNEZ-VILA, E., SERENA, J., SEGURA, T., CRUZ, V. T. & MASJUAN, J. 2012. Citicoline in the treatment of acute ischaemic stroke: an international, randomised, multicentre, placebo-controlled study (ICTUS trial). The Lancet, 380, 349-357.
47. DE YEBENES, J. G., LANDWEHRMEYER, B., SQUITIERI, F., REILMANN, R., ROSSER, A., BARKER, R. A., SAFT, C., MAGNET, M. K., SWORD, A. & REMBRATT, Å. 2011. Pridopidine for the treatment of motor function in patients with Huntington's disease (MermaiHD): a phase 3, randomised, double-blind, placebo-controlled trial. The Lancet Neurology, 10, 1049-1057.
48. DEVANAND, D., MINTZER, J., SCHULTZ, S. K., ANDREWS, H. F., SULTZER, D. L., DE LA PENA, D., GUPTA, S., COLON, S., SCHIMMING, C. & PELTON, G. H. 2012. Relapse risk after discontinuation of risperidone in Alzheimer's disease. New England Journal of Medicine, 367, 1497-1507.
49. DLUGOS, D., SHINNAR, S., CNAAN, A., HU, F., MOSHÉ, S., MIZRAHI, E., MASUR, D., SOGAWA, Y., LE PICHON, J. & LEVINE, C. 2013. Pretreatment EEG in childhood absence epilepsy Associations with attention and treatment outcome. Neurology, 81, 150-156.
50. DOBKIN, R. D., MENZA, M., ALLEN, L. A., GARA, M. A., MARK, M. H., TIU, J., BIENFAIT, K. L. & FRIEDMAN, J. 2011. Cognitive behavior therapy for depression in Parkinson's disease: A randomized controlled trial. The American journal of psychiatry, 168, 1066.
51. DOODY, R. S., RAMAN, R., FARLOW, M., IWATSUBO, T., VELLAS, B., JOFFE, S., KIEBURTZ, K., HE, F., SUN, X. & THOMAS, R. G. 2013. A phase 3 trial of semagacestat for treatment of Alzheimer's disease. New England Journal of Medicine, 369, 341-350.
52. DOODY, R. S., THOMAS, R. G., FARLOW, M., IWATSUBO, T., VELLAS, B., JOFFE, S., KIEBURTZ, K., RAMAN, R., SUN, X. & AISEN, P. S. 2014. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. New England Journal of Medicine, 370, 311-321.
53. DRAKE, R. E., FREY, W., BOND, G. R., GOLDMAN, H. H., SALKEVER, D., MILLER, A., MOORE, T. A., RILEY, J., KARAKUS, M. & MILFORT, R. 2013. Assisting Social Security Disability Insurance beneficiaries with schizophrenia, bipolar disorder, or major depression in returning to work. American Journal of Psychiatry, 170, 1433-1441.
54. DRIESSEN, E., VAN, H. L., DON, F. J., PEEN, J., KOOL, S., WESTRA, D., HENDRIKSEN, M., SCHOEVERS, R. A., CUIJPERS, P. & TWISK, J. W. 2013. The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: a randomized clinical trial. American Journal of Psychiatry, 170, 1041-1050.

55. DRUSS, B. G., JI, X., GLICK, G. & SILKE, A. 2014. Randomized trial of an electronic personal health record for patients with serious mental illnesses. American Journal of Psychiatry, 171, 360-368.
56. EDERLE, J., DAVAGNANAM, I., VAN DER WORP, H. B., VENABLES, G. S., LYRER, P. A., FEATHERSTONE, R. L., BROWN, M. M. & JÄGER, H. R. 2013. Effect of white-matter lesions on the risk of periprocedural stroke after carotid artery stenting versus endarterectomy in the International Carotid Stenting Study (ICSS): a prespecified analysis of data from a randomised trial. The Lancet Neurology, 12, 866-872.
57. EDWARDS, J. D., HAUSER, R. A., O'CONNOR, M. L., VALDÉS, E. G., ZESIEWICZ, T. A. & UC, E. Y. 2013. Randomized trial of cognitive speed of processing training in Parkinson disease. Neurology, 81, 1284-1290.
58. EHLERS, A., HACKMANN, A., GREY, N., WILD, J., LINESS, S., ALBERT, I., DEALE, A., STOTT, R. & CLARK, D. M. 2014. A randomized controlled trial of 7-day intensive and standard weekly cognitive therapy for PTSD and emotion-focused supportive therapy. American Journal of Psychiatry, 171, 294-304.
59. ESCOLAR, D., HACHE, L., CLEMENS, P., CNAAN, A., MCDONALD, C., VISWANATHAN, V., KORNBERG, A., BERTORINI, T., NEVO, Y. & LOTZE, T. 2011. Randomized, blinded trial of weekend vs daily prednisone in Duchenne muscular dystrophy. Neurology, 77, 444-452.
60. ESCOLAR, D., ZIMMERMAN, A., BERTORINI, T., CLEMENS, P., CONNOLLY, A., MESA, L., GORNI, K., KORNBERG, A., KOLSKI, H. & KUNTZ, N. 2012. Pentoxifylline as a rescue treatment for DMD A randomized double-blind clinical trial. Neurology, 78, 904-913.
61. ESSOCK, S. M., SCHOOLER, N. R., STROUP, T. S., MCEVOY, J. P., ROJAS, I., JACKSON, C. & COVELL, N. H. 2011. Effectiveness of switching from antipsychotic polypharmacy to monotherapy. American Journal of Psychiatry, 168, 702-708.
62. FEDER, G., DAVIES, R. A., BAIRD, K., DUNNE, D., ELDRIDGE, S., GRIFFITHS, C., GREGORY, A., HOWELL, A., JOHNSON, M. & RAMSAY, J. 2011. Identification and Referral to Improve Safety (IRIS) of women experiencing domestic violence with a primary care training and support programme: a cluster randomised controlled trial. The Lancet, 378, 1788-1795.
63. FORSTER, A., DICKERSON, J., YOUNG, J., PATEL, A., KALRA, L., NIXON, J., SMITHARD, D., KNAPP, M., HOLLOWAY, I. & ANWAR, S. 2014. A structured training programme for caregivers of inpatients after stroke (TRACS): a cluster randomised controlled trial and cost-effectiveness analysis. The Lancet, 382, 2069-2076.
64. FORTNEY, J. C., PYNE, J. M., MOUDEN, S. B., MITTAL, D., HUDSON, T. J., SCHROEDER, G. W., WILLIAMS, D. K., BYNUM, C. A., MATTOX, R. & ROST, K. M. 2013. Practice-based versus telemedicine-based collaborative care for depression in rural federally qualified health centers: a pragmatic randomized comparative effectiveness trial. American Journal of Psychiatry, 170, 414-425.
65. FOX, R. J., MILLER, D. H., PHILLIPS, J. T., HUTCHINSON, M., HAVRDOVA, E., KITA, M., YANG, M., RAGHUPATHI, K., NOVAS, M. & SWEETSER, M. T. 2012. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. New England Journal of Medicine, 367, 1087-1097.
66. FREE, C., KNIGHT, R., ROBERTSON, S., WHITTAKER, R., EDWARDS, P., ZHOU, W., RODGERS, A., CAIRNS, J., KENWARD, M. G. & ROBERTS, I. 2011. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. The Lancet, 378, 49-55.
67. FREEDMAN, M., BAR-OR, A., OGER, J., TRABOULSEE, A., PATRY, D., YOUNG, C., OLSSON, T., LI, D., HARTUNG, H.-P. & KRANTZ, M. 2011. A phase III study evaluating the efficacy and safety of MBP8298 in secondary progressive MS. Neurology, 77, 1551-1560.
68. FREEMAN, J., SAPYTA, J., GARCIA, A., COMPTON, S., KHANNA, M., FLESSNER, C., FITZGERALD, D., MAURO, C., DINGFELDER, R. & BENITO, K. 2014. Family-Based Treatment of Early Childhood Obsessive-Compulsive Disorder: The Pediatric Obsessive-Compulsive Disorder Treatment Study for Young Children (POTS Jr)-A Randomized Clinical Trial. JAMA psychiatry.

69. FRENCH, J. A., KRAUSS, G. L., BITON, V., SQUILLACOTE, D., YANG, H., LAURENZA, A., KUMAR, D., ROGAWSKI, M. A., CAMPANILLE, V. & FLORIDIA, J. 2012. Adjunctive perampanel for refractory partial-onset seizures Randomized phase III study 304. Neurology, 79, 589-596.
70. FRIEDMAN, B. W., GARBER, L., YOON, A., SOLORZANO, C., WOLLOWITZ, A., ESSES, D., BIJUR, P. E. & GALLAGHER, E. J. 2014. Randomized trial of IV valproate vs metoclopramide vs ketorolac for acute migraine. Neurology, 82, 976-983.
71. FRITZ, D., BROUWER, M. C. & VAN DE BEEK, D. 2012. Dexamethasone and long-term survival in bacterial meningitis. Neurology, 79, 2177-2179.
72. GELLER, B., LUBY, J. L., JOSHI, P., WAGNER, K. D., EMSLIE, G., WALKUP, J. T., AXELSON, D. A., BOLHOFNER, K., ROBB, A. & WOLF, D. V. 2012. A randomized controlled trial of risperidone, lithium, or divalproex sodium for initial treatment of bipolar I disorder, manic or mixed phase, in children and adolescents. Archives of general psychiatry, archgenpsychiatry. 2011.1508 v1.
73. GIACINO, J. T., WHYTE, J., BAGIELLA, E., KALMAR, K., CHILDS, N., KHADEMI, A., EIFERT, B., LONG, D., KATZ, D. I. & CHO, S. 2012. Placebo-controlled trial of amantadine for severe traumatic brain injury. New England Journal of Medicine, 366, 819-826.
74. GIESEN-BLOO, J., VAN DYCK, R., SPINHOVEN, P., VAN TILBURG, W., DIRKSEN, C., VAN ASSELT, T., KREMERS, I., NADORT, M. & ARNTZ, A. 2006. Outpatient psychotherapy for borderline personality disorder: randomized trial of schema-focused therapy vs transference-focused psychotherapy. Archives of general psychiatry, 63, 649-658.
75. GINSBERG, M. D., PALESCH, Y. Y., HILL, M. D., MARTIN, R. H., MOY, C. S., BARSAN, W. G., WALDMAN, B. D., TAMARIZ, D. & RYCKBORST, K. J. 2013. High-dose albumin treatment for acute ischaemic stroke (ALIAS) part 2: a randomised, double-blind, phase 3, placebo-controlled trial. The Lancet Neurology, 12, 1049-1058.
76. GOLD, R., GIOVANNONI, G., SELMAJ, K., HAVRDOVA, E., MONTALBAN, X., RADUE, E.-W., STEFOSKI, D., ROBINSON, R., RIESTER, K. & RANA, J. 2013. Daclizumab high-yield process in relapsing-remitting multiple sclerosis (SELECT): a randomised, double-blind, placebo-controlled trial. The Lancet, 381, 2167-2175.
77. GOLD, R., KAPPOS, L., ARNOLD, D. L., BAR-OR, A., GIOVANNONI, G., SELMAJ, K., TORNATORE, C., SWEETSER, M. T., YANG, M. & SHEIKH, S. I. 2012. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. New England Journal of Medicine, 367, 1098-1107.
78. GOLDIN, P. R., ZIV, M., JAZAIERI, H., HAHN, K., HEIMBERG, R. & GROSS, J. J. 2013. Impact of cognitive behavioral therapy for social anxiety disorder on the neural dynamics of cognitive reappraisal of negative self-beliefs: randomized clinical trial. JAMA psychiatry, 70, 1048-1056.
79. GRANT, P. M., HUH, G. A., PERIVOLIOTIS, D., STOLAR, N. M. & BECK, A. T. 2012. Randomized trial to evaluate the efficacy of cognitive therapy for low-functioning patients with schizophrenia. Archives of general psychiatry, 69, 121.
80. GRAY, K. M., CARPENTER, M. J., BAKER, N. L., DESANTIS, S. M., KRYWAY, E., HARTWELL, K. J., MCRAE-CLARK, A. L. & BRADY, K. T. 2012. A double-blind randomized controlled trial of N-acetylcysteine in cannabis-dependent adolescents. American Journal of Psychiatry, 169, 805-812.
81. GUSTAFSON, D. H., MCTAVISH, F. M., CHIH, M.-Y., ATWOOD, A. K., JOHNSON, R. A., BOYLE, M. G., LEVY, M. S., DRISCOLL, H., CHISHOLM, S. M. & DILLENBURG, L. 2014. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA psychiatry.
82. HANNEY, M., PRASHER, V., WILLIAMS, N., JONES, E. L., AARSLAND, D., CORBETT, A., LAWRENCE, D., YU, L.-M., TYRER, S. & FRANCIS, P. T. 2012. Memantine for dementia in adults older than 40 years with Down's syndrome (MEADOWS): a randomised, double-blind, placebo-controlled trial. The Lancet, 379, 528-536.

83. HATTA, K., KISHI, Y., WADA, K., TAKEUCHI, T., ODAWARA, T., USUI, C. & NAKAMURA, H. 2014. Preventive effects of ramelteon on delirium: a randomized placebo-controlled trial. JAMA psychiatry, 71, 397-403.
84. HAUSER, R. A., HSU, A., KELL, S., ESPAY, A. J., SETHI, K., STACY, M., ONDO, W., O'CONNELL, M. & GUPTA, S. 2013. Extended-release carbidopa-levodopa (IPX066) compared with immediate-release carbidopa-levodopa in patients with Parkinson's disease and motor fluctuations: a phase 3 randomised, double-blind trial. The Lancet Neurology.
85. HENDERSON, V., JOHN, J. S., HODIS, H., KONO, N., MCCLEARY, C., FRANKE, A. & MACK, W. 2012. Long-term soy isoflavone supplementation and cognition in women A randomized, controlled trial. Neurology, 78, 1841-1848.
86. HERZOG, A. G., FOWLER, K. M., SMITHSON, S. D., KALAYJIAN, L. A., HECK, C. N., SPERLING, M. R., LIPORACE, J. D., HARDEN, C. L., DWORETZKY, B. A., PENNELL, P. B. & MASSARO, J. M. 2012. Progesterone vs placebo therapy for women with epilepsy: A randomized clinical trial. Neurology, 78, 1959-66.
87. HOFMANN, S. G., SMITS, J. A., ROSENFIELD, D., SIMON, N., OTTO, M. W., MEURET, A. E., MARQUES, L., FANG, A., TART, C. & POLLACK, M. H. 2013. D-Cycloserine as an augmentation strategy with cognitive-behavioral therapy for social anxiety disorder. American Journal of Psychiatry, 170, 751-758.
88. HOLLANDER, E., SOORYA, L., CHAPLIN, W., ANAGNOSTOU, E., TAYLOR, B. P., FERRETTI, C. J., WASSERMAN, S., SWANSON, E. & SETTIPANI, C. 2012. A double-blind placebo-controlled trial of fluoxetine for repetitive behaviors and global severity in adult autism spectrum disorders. American Journal of Psychiatry, 169, 292-299.
89. HONG, L. E., THAKER, G. K., MCMAHON, R. P., SUMMERFELT, A., RACHBEISEL, J., FULLER, R. L., WONODI, I., BUCHANAN, R. W., MYERS, C. & HEISHMAN, S. J. 2011. Effects of moderate-dose treatment with varenicline on neurobiological and cognitive biomarkers in smokers and nonsmokers with schizophrenia or schizoaffective disorder. Archives of general psychiatry, 68, 1195-1206.
90. HOWARD, R., MCSHANE, R., LINDESAY, J., RITCHIE, C., BALDWIN, A., BARBER, R., BURNS, A., DENING, T., FINDLAY, D. & HOLMES, C. 2012. Donepezil and memantine for moderate-to-severe Alzheimer's disease. New England Journal of Medicine, 366, 893-903.
91. JARRETT, R. B., MINHAJUDDIN, A., GERSHENFELD, H., FRIEDMAN, E. S. & THASE, M. E. 2013. Preventing depressive relapse and recurrence in higher-risk cognitive therapy responders: a randomized trial of continuation phase cognitive therapy, fluoxetine, or matched pill placebo. JAMA psychiatry, 70, 1152-1160.
92. JARSKOG, L. F., HAMER, R. M., CATELLIER, D. J., STEWART, D. D., LAVANGE, L., RAY, N., GOLDEN, L. H., LIEBERMAN, J. A. & STROUP, T. S. 2013. Metformin for weight loss and metabolic control in overweight outpatients with schizophrenia and schizoaffective disorder. American Journal of Psychiatry, 170, 1032-1040.
93. JOHNSON, B. A., AIT-DAOUD, N., WANG, X.-Q., PENBERTHY, J. K., JAVORS, M. A., SENEVIRATNE, C. & LIU, L. 2013. Topiramate for the treatment of cocaine addiction: a randomized clinical trial. JAMA psychiatry, 70, 1338-1346.
94. JÜTTLER, E., UNTERBERG, A., WOITZIK, J., BÖSEL, J., AMIRI, H., SAKOWITZ, O. W., GONDAN, M., SCHILLER, P., LIMPRECHT, R. & LUNTZ, S. 2014. Hemicraniectomy in older patients with extensive middle-cerebral-artery stroke. New England Journal of Medicine, 370, 1091-1100.
95. KATON, W., RUSSO, J., LIN, E. H., SCHMITTDIEL, J., CIECHANOWSKI, P., LUDMAN, E., PETERSON, D., YOUNG, B. & VON KORFF, M. 2012. Cost-effectiveness of a multicondition collaborative care intervention: a randomized controlled trial. Archives of general psychiatry, 69, 506-514.
96. KIDWELL, C. S., JAHAN, R., GORNBEIN, J., ALGER, J. R., NENOV, V., AJANI, Z., FENG, L., MEYER, B. C., OLSON, S. & SCHWAMM, L. H. 2013. A trial of imaging selection and endovascular treatment for ischemic stroke. New England Journal of Medicine, 368, 914-923.

97. KIM, J.-S., OH, S.-Y., LEE, S.-H., KANG, J.-H., KIM, D., JEONG, S.-H., CHOI, K.-D., MOON, I.-S., KIM, B.-K. & OH, H. 2012a. Randomized clinical trial for apogeotropic horizontal canal benign paroxysmal positional vertigo. Neurology, 78, 159-166.
98. KIM, J. S., OH, S.-Y., LEE, S.-H., KANG, J. H., KIM, D. U., JEONG, S.-H., CHOI, K.-D., MOON, I. S., KIM, B. K. & KIM, H. J. 2012b. Randomized clinical trial for geotropic horizontal canal benign paroxysmal positional vertigo. Neurology, 79, 700-707.
99. KIM, S., KIM, H., KNOPMAN, D., DE VRIES, R., DAMSCHRODER, L. & APPELBAUM, P. 2011. Effect of public deliberation on attitudes toward surrogate consent for dementia research. Neurology, 77, 2097-2104.
100. KLAMROTH-MARGANSKA, V., BLANCO, J., CAMPEN, K., CURT, A., DIETZ, V., ETTLIN, T., FELDER, M., FELLINGHAUER, B., GUIDALI, M. & KOLLMAR, A. 2014. Three-dimensional, task-specific robot therapy of the arm after stroke: a multicentre, parallel-group randomised trial. The Lancet Neurology, 13, 159-166.
101. KOPELOWICZ, A., ZARATE, R., WALLACE, C. J., LIBERMAN, R. P., LOPEZ, S. R. & MINTZ, J. 2012. The ability of multifamily groups to improve treatment adherence in Mexican Americans with schizophrenia. Archives of general psychiatry, 69, 265-273.
102. KRANZLER, H. R., COVAULT, J., FEINN, R., ARMELI, S., TENNEN, H., ARIAS, A. J., GELERNTER, J., POND, T., ONCKEN, C. & KAMPMAN, K. M. 2014. Topiramate treatment for heavy drinkers: moderation by a GRIK1 polymorphism. American Journal of Psychiatry, 171, 445-452.
103. KRAUSS, G., SERRATOSA, J., VILLANUEVA, V., ENDZINIENE, M., HONG, Z., FRENCH, J., YANG, H., SQUILLACOTE, D., EDWARDS, H. & ZHU, J. 2012. Randomized phase III study 306 Adjunctive perampanel for refractory partial-onset seizures. Neurology, 78, 1408-1415.
104. KRUPITSKY, E., ZVARTAU, E., BLOKHINA, E., VERBITSKAYA, E., WAHLGREN, V., TSOY-PODOSENIN, M., BUSHARA, N., BURAKOV, A., MASALOV, D. & ROMANOVA, T. 2012. Randomized trial of long-acting sustained-release naltrexone implant vs oral naltrexone or placebo for preventing relapse to opioid dependence. Archives of general psychiatry, 69, 973.
105. KWAN, P., BRODIE, M. J., KÄLVIÄINEN, R., YURKEWICZ, L., WEAVER, J. & KNAPP, L. E. 2011. Efficacy and safety of pregabalin versus lamotrigine in patients with newly diagnosed partial seizures: a phase 3, double-blind, randomised, parallel-group trial. The Lancet Neurology, 10, 881-890.
106. LANE, H.-Y., LIN, C.-H., GREEN, M. F., HELLEMANN, G., HUANG, C.-C., CHEN, P.-W., TUN, R., CHANG, Y.-C. & TSAI, G. E. 2013. Add-on treatment of benzoate for schizophrenia: a randomized, double-blind, placebo-controlled trial of d-amino acid oxidase inhibitor. JAMA psychiatry, 70, 1267-1275.
107. LAUNER, L. J., MILLER, M. E., WILLIAMSON, J. D., LAZAR, R. M., GERSTEIN, H. C., MURRAY, A. M., SULLIVAN, M., HOROWITZ, K. R., DING, J. & MARCOVINA, S. 2011. Effects of intensive glucose lowering on brain structure and function in people with type 2 diabetes (ACCORD MIND): a randomised open-label substudy. The Lancet Neurology, 10, 969-977.
108. LÉGER, J.-M., VIALA, K., NICOLAS, G., CRÉANGE, A., VALLAT, J.-M., POUGET, J., CLAVELOU, P., VIAL, C., STECK, A. & MUSSET, L. 2013. Placebo-controlled trial of rituximab in IgM anti-myelin–associated glycoprotein neuropathy. Neurology, 80, 2217-2225.
109. LEICHSENRING, F., SALZER, S., BEUTEL, M. E., HERPERTZ, S., HILLER, W., HOYER, J., HUESING, J., JORASCHKY, P., NOLTING, B. & POEHLMANN, K. 2013. Psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder: a multicenter randomized controlled trial. American Journal of Psychiatry, 170, 759-767.
110. LEIST, T. P., COMI, G., CREE, B. A., COYLE, P. K., FREEDMAN, M. S., HARTUNG, H.-P., VERMERSCH, P., CASSET-SEMANAZ, F. & SCARAMOZZA, M. 2014. Effect of oral cladribine on time to conversion to clinically definite multiple sclerosis in patients with a first demyelinating event (ORACLE MS): a phase 3 randomised trial. The Lancet Neurology, 13, 257-267.

111. LEONTJEVAS, R., GERRITSEN, D. L., SMALBRUGGE, M., TEERENSTRA, S., VERNOOIJ-DASSEN, M. J. & KOOPMANS, R. T. 2013. A structural multidisciplinary approach to depression management in nursing-home residents: a multicentre, stepped-wedge cluster-randomised trial. The Lancet, 381, 2255-2264.
112. LEWITT, P. A., HAUSER, R. A., LU, M., NICHOLAS, A. P., WEINER, W., COPPARD, N., LEINONEN, M. & SAVOLA, J.-M. 2012. Randomized clinical trial of fipamezole for dyskinesia in Parkinson disease (FJORD study). Neurology, 79, 163-169.
113. LI, F., HARMER, P., FITZGERALD, K., ECKSTROM, E., STOCK, R., GALVER, J., MADDALOZZO, G. & BATYA, S. S. 2012. Tai chi and postural stability in patients with Parkinson's disease. New England Journal of Medicine, 366, 511-519.
114. LOEBEL, A., CUCCHIARO, J., SILVA, R., KROGER, H., HSU, J., SARMA, K. & SACHS, G. 2014a. Lurasidone monotherapy in the treatment of bipolar I depression: a randomized, double-blind, placebo-controlled study. American Journal of Psychiatry, 171, 160-168.
115. LOEBEL, A., CUCCHIARO, J., SILVA, R., KROGER, H., SARMA, K., XU, J. & CALABRESE, J. R. 2014b. Lurasidone as adjunctive therapy with lithium or valproate for the treatment of bipolar I depression: a randomized, double-blind, placebo-controlled study. American Journal of Psychiatry, 171, 169-177.
116. LOVERA, J. F., KIM, E., HERIZA, E., FITZPATRICK, M., HUNZIKER, J., TURNER, A. P., ADAMS, J., STOVER, T., SANGEORZAN, A. & SLOAN, A. 2012. Ginkgo biloba does not improve cognitive function in MS A randomized placebo-controlled trial. Neurology, 79, 1278-1284.
117. MACDONALD, R. L., HIGASHIDA, R. T., KELLER, E., MAYER, S. A., MOLYNEUX, A., RAABE, A., VAJKOCZY, P., WANKE, I., BACH, D. & FREY, A. 2011. Clazosentan, an endothelin receptor antagonist, in patients with aneurysmal subarachnoid haemorrhage undergoing surgical clipping: a randomised, double-blind, placebo-controlled phase 3 trial (CONSCIOUS-2). The Lancet Neurology, 10, 618-625.
118. MANCONI, M., FERRI, R., ZUCCONI, M., CLEMENS, S., GIAROLLI, L., BOTTASINI, V. & FERINI-STRAMBI, L. 2011. Preferential D2 or preferential D3 dopamine agonists in restless legs syndrome. Neurology, 77, 110-117.
119. MARSHALL, R. S., FESTA, J. R., CHEUNG, Y.-K., PAVOL, M. A., DERDEYN, C. P., CLARKE, W. R., VIDEEN, T. O., GRUBB, R. L., SLANE, K. & POWERS, W. J. 2014. Randomized Evaluation of Carotid Occlusion and Neurocognition (RECON) trial Main results. Neurology, 82, 744-751.
120. MASUR, D., SHINNAR, S., CNAAN, A., SHINNAR, R. C., CLARK, P., WANG, J., WEISS, E. F., HIRTZ, D. G. & GLAUSER, T. A. 2013. Pretreatment cognitive deficits and intervention effects on attention in childhood absence epilepsy. Neurology, 81, 1572-1580.
121. MCDONELL, M. G., SREBNIK, D., ANGELO, F., MCPHERSON, S., LOWE, J. M., SUGAR, A., SHORT, R. A., ROLL, J. M. & RIES, R. K. 2013. Randomized controlled trial of contingency management for stimulant use in community mental health patients with serious mental illness. American Journal of Psychiatry, 170, 94-101.
122. MCGRATH, C. L., KELLEY, M. E., HOLTZHEIMER, P. E., DUNLOP, B. W., CRAIGHEAD, W. E., FRANCO, A. R., CRADDOCK, R. C. & MAYBERG, H. S. 2013. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA psychiatry, 70, 821-829.
123. MEES, S. M. D., ALGRA, A., VANDERTOP, W. P., VAN KOOTEN, F., KUIJSTEN, H. A., BOITEN, J., VAN OOSTENBRUGGE, R. J., SALMAN, R. A.-S., LAVADOS, P. M. & RINKEL, G. J. 2012. Magnesium for aneurysmal subarachnoid haemorrhage (MASH-2): a randomised placebo-controlled trial. The Lancet, 380, 44-49.
124. MEIER, B., KALESAN, B., MATTLE, H. P., KHATTAB, A. A., HILDICK-SMITH, D., DUDEK, D., ANDERSEN, G., IBRAHIM, R., SCHULER, G. & WALTON, A. S. 2013. Percutaneous closure of patent foramen ovale in cryptogenic embolism. New England Journal of Medicine, 368, 1083-1091.
125. MELTZER, H. Y., CUCCHIARO, J., SILVA, R., OGASA, M., PHILLIPS, D., XU, J., KALALI, A. H., SCHWEIZER, E., PIKALOV, A. & LOEBEL, A. 2011. Lurasidone in the treatment of

schizophrenia: a randomized, double-blind, placebo-and olanzapine-controlled study. American Journal of Psychiatry, 168, 957-967.

126. MENG, R., ASMARO, K., MENG, L., LIU, Y., MA, C., XI, C., LI, G., REN, C., LUO, Y. & LING, F. 2012. Upper limb ischemic preconditioning prevents recurrent stroke in intracranial arterial stenosis. Neurology, 79, 1853-1861.
127. MICHELSON, D., SNYDER, E., PARADIS, E., CHENGAN-LIU, M., SNAVELY, D. B., HUTZELMANN, J., WALSH, J. K., KRYSTAL, A. D., BENCA, R. M. & COHN, M. 2014. Safety and efficacy of suvorexant during 1-year treatment of insomnia with subsequent abrupt treatment discontinuation: a phase 3 randomised, double-blind, placebo-controlled trial. The Lancet Neurology, 13, 461-471.
128. MIDDLETON, S., MCELDUFF, P., WARD, J., GRIMSHAW, J. M., DALE, S., D'ESTE, C., DRURY, P., GRIFFITHS, R., CHEUNG, N. W. & QUINN, C. 2011. Implementation of evidence-based treatment protocols to manage fever, hyperglycaemia, and swallowing dysfunction in acute stroke (QASC): a cluster randomised controlled trial. The Lancet, 378, 1699-1706.
129. MIKLOWITZ, D. J., SCHNECK, C. D., GEORGE, E. L., TAYLOR, D. O., SUGAR, C. A., BIRMAHER, B., KOWATCH, R. A., DELBELLO, M. P. & AXELSON, D. A. 2014. Pharmacotherapy and family-focused treatment for adolescents with bipolar I and II disorders: a 2-year randomized trial. American Journal of Psychiatry, 171, 658-667.
130. MOREAU, C., DELVAL, A., DEFEBVRE, L., DUJARDIN, K., DUHAMEL, A., PETYT, G., VUILLAUME, I., CORVOL, J.-C., BREFEL-COURBON, C. & ORY-MAGNE, F. 2012. Methylphenidate for gait hypokinesia and freezing in patients with Parkinson's disease undergoing subthalamic stimulation: a multicentre, parallel, randomised, placebo-controlled trial. The Lancet Neurology, 11, 589-596.
131. MURROUGH, J. W., IOSIFESCU, D. V., CHANG, L. C., AL JURDI, R. K., GREEN, C. E., PEREZ, A. M., IQBAL, S., PILLEMER, S., FOULKES, A. & SHAH, A. 2013. Antidepressant efficacy of ketamine in treatment-resistant major depression: a two-site randomized controlled trial. American Journal of Psychiatry, 170, 1134-1142.
132. NAKASUJJA, N., MIYAHARA, S., EVANS, S., LEE, A., MUSISI, S., KATABIRA, E., ROBERTSON, K., RONALD, A., CLIFFORD, D. B. & SACKTOR, N. 2013. Randomized trial of minocycline in the treatment of HIV-associated cognitive impairment. Neurology, 80, 196-202.
133. NG, Y., CONRY, J., DRUMMOND, R., STOLLE, J. & WEINBERG, M. 2011. Randomized, phase III study results of clobazam in Lennox-Gastaut syndrome. Neurology, 77, 1473-1481.
134. NIERENBERG, A. A., FRIEDMAN, E. S., BOWDEN, C. L., SYLVIA, L. G., THASE, M. E., KETTER, T., OSTACHER, M. J., LEON, A. C., REILLY-HARRINGTON, N. & IOSIFESCU, D. V. 2013. Lithium Treatment Moderate-Dose Use Study (LiTMUS) for bipolar disorder: a randomized comparative effectiveness trial of optimized personalized treatment with and without lithium. American Journal of Psychiatry, 170, 102-110.
135. NOBILE-ORAZIO, E., COCITO, D., JANN, S., UNCINI, A., BEGHI, E., MESSINA, P., ANTONINI, G., FAZIO, R., GALLIA, F. & SCHENONE, A. 2012. Intravenous immunoglobulin versus intravenous methylprednisolone for chronic inflammatory demyelinating polyradiculoneuropathy: a randomised controlled trial. The Lancet Neurology, 11, 493-502.
136. NOGUEIRA, R. G., LUTSEP, H. L., GUPTA, R., JOVIN, T. G., ALBERS, G. W., WALKER, G. A., LIEBESKIND, D. S. & SMITH, W. S. 2012. Trevo versus Merci retrievers for thrombectomy revascularisation of large vessel occlusions in acute ischaemic stroke (TREVO 2): a randomised trial. The Lancet, 380, 1231-1240.
137. O'CONNOR, P., WOLINSKY, J. S., CONFAVREUX, C., COMI, G., KAPPOS, L., OLSSON, T. P., BENZERDJEB, H., TRUFFINET, P., WANG, L. & MILLER, A. 2011. Randomized trial of oral teriflunomide for relapsing multiple sclerosis. New England Journal of Medicine, 365, 1293-1303.
138. ODEKERKEN, V. J., VAN LAAR, T., STAAL, M. J., MOSCH, A., HOFFMANN, C. F., NIJSSEN, P. C., BEUTE, G. N., VAN VUGT, J. P., LENDERS, M. W. & CONTARINO, M. F. 2013. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. The Lancet Neurology, 12, 37-44.

139. OKUN, M. S., GALLO, B. V., MANDYBUR, G., JAGID, J., FOOTE, K. D., REVILLA, F. J., ALTERMAN, R., JANKOVIC, J., SIMPSON, R. & JUNN, F. 2012. Subthalamic deep brain stimulation with a constant-current device in Parkinson's disease: an open-label randomised controlled trial. The Lancet Neurology, 11, 140-149.
140. OLANOW, C. W., KIEBURTZ, K., ODIN, P., ESPAY, A. J., STANDAERT, D. G., FERNANDEZ, H. H., VANAGUNAS, A., OTHMAN, A. A., WIDNELL, K. L. & ROBIESON, W. Z. 2014. Continuous intrajejunal infusion of levodopa-carbidopa intestinal gel for patients with advanced Parkinson's disease: a randomised, controlled, double-blind, double-dummy study. The Lancet Neurology, 13, 141-149.
141. ONDO, W., KENNEY, C., SULLIVAN, K., DAVIDSON, A., HUNTER, C., JAHAN, I., MCCOMBS, A., MILLER, A. & ZESIEWICZ, T. 2012. Placebo-controlled trial of lubiprostone for constipation associated with Parkinson disease. Neurology, 78, 1650-1654.
142. OQUENDO, M. A., GALFALVY, H. C., CURRIER, D., GRUNEBAUM, M. F., SHER, L., SULLIVAN, G. M., BURKE, A. K., HARKAVY-FRIEDMAN, J., SUBLETTE, M. E. & PARSEY, R. V. 2011. Treatment of suicide attempters with bipolar disorder: a randomized clinical trial comparing lithium and valproate in the prevention of suicidal behavior. American journal of psychiatry, 168, 1050-1056.
143. ORY-MAGNE, F., CORVOL, J.-C., AZULAY, J.-P., BONNET, A.-M., BREFEL-COURBON, C., DAMIER, P., DELLAPINA, E., DESTÉE, A., DURIF, F. & GALITZKY, M. 2014. Withdrawing amantadine in dyskinetic patients with Parkinson disease The AMANDYSK trial. Neurology, 82, 300-307.
144. OSYPUK, T. L., TCHETGEN, E. J. T., ACEVEDO-GARCIA, D., EARLS, F. J., LINCOLN, A., SCHMIDT, N. M. & GLYMOUR, M. M. 2012. Differential mental health effects of neighborhood relocation among youth in vulnerable families: results from a randomized trial. Archives of general psychiatry, 69, 1284-1294.
145. PAPAKOSTAS, G. I., SHELTON, R. C., ZAJECKA, J. M., ETEMAD, B., RICKELS, K., CLAIN, A., BAER, L., DALTON, E. D., SACCO, G. R. & SCHOENFELD, D. 2012. L-methylfolate as adjunctive therapy for SSRI-resistant major depression: results of two randomized, double-blind, parallel-sequential trials. American Journal of Psychiatry, 169, 1267-1274.
146. PLEWNIA, C., VONTHEIN, R., WASSERKA, B., ARFELLER, C., NAUMANN, A., SCHRAVEN, S. & PLONTKE, S. 2012. Treatment of chronic tinnitus with theta burst stimulation A randomized controlled trial. Neurology, 78, 1628-1634.
147. POEWE, W., RASCOL, O., BARONE, P., HAUSER, R., MIZUNO, Y., HAAKSMA, M., SALIN, L., JUHEL, N. & SCHAPIRA, A. 2011. Extended-release pramipexole in early Parkinson disease A 33-week randomized controlled trial. Neurology, 77, 759-766.
148. POLLACK, M. H., VAN AMERINGEN, M., SIMON, N. M., WORTHINGTON, J. W., HOGE, E. A., KESHAVIAH, A. & STEIN, M. B. 2014. A double-blind randomized controlled trial of augmentation and switch strategies for refractory social anxiety disorder. American Journal of Psychiatry, 171, 44-53.
149. POSTUMA, R. B., LANG, A. E., MUNHOZ, R. P., CHARLAND, K., PELLETIER, A., MOSCOVICH, M., FILLA, L., ZANATTA, D., ROMENETS, S. R. & ALTMAN, R. 2012. Caffeine for treatment of Parkinson disease A randomized controlled trial. Neurology, 79, 651-658.
150. POULSEN, S., LUNN, S., DANIEL, S. I., FOLKE, S., MATHIESEN, B. B., KATZNELSON, H. & FAIRBURN, C. G. 2014. A randomized controlled trial of psychoanalytic psychotherapy or cognitive-behavioral therapy for bulimia nervosa. American Journal of Psychiatry, 171, 109-116.
151. RAMOS, V. F. M. L., PAINE, R. W., THIRUGNANASAMBANDAM, N., SHIROTA, Y., HAMADA, M. & UGAWA, Y. 2013. Supplementary motor area stimulation for Parkinson disease: A randomized controlled study. Neurology, 81, 1881-1882.
152. RASKIND, M. A., PETERSON, K., WILLIAMS, T., HOFF, D. J., HART, K., HOLMES, H., HOMAS, D., HILL, J., DANIELS, C. & CALOHAN, J. 2013. A trial of prazosin for combat trauma PTSD with nightmares in active-duty soldiers returned from Iraq and Afghanistan. American Journal of Psychiatry, 170, 1003-1010.

153. RAZ, L., JAYACHANDRAN, M., TOSAKULWONG, N., LESNICK, T. G., WILLE, S. M., MURPHY, M. C., SENJEM, M. L., GUNTER, J. L., VEMURI, P. & JACK, C. R. 2013. Thrombogenic microvesicles and white matter hyperintensities in postmenopausal women. Neurology, 80, 911-918.
154. RICHARD, I., MCDERMOTT, M., KURLAN, R., LYNESS, J., COMO, P., PEARSON, N., FACTOR, S., JUNCOS, J., RAMOS, C. S. & BRODSKY, M. 2012. A randomized, double-blind, placebo-controlled trial of antidepressants in Parkinson disease. Neurology, 78, 1229-1236.
155. RISTORI, G., ROMANO, S., CANNONI, S., VISCONTI, A., TINELLI, E., MENDOZZI, L., CECCONI, P., LANZILLO, R., QUARANTELLI, M. & BUTTINELLI, C. 2014. Effects of Bacille Calmette-Guérin after the first demyelinating event in the CNS. Neurology, 82, 41-48.
156. ROFFMAN, J. L., LAMBERTI, J. S., ACHTYES, E., MACKLIN, E. A., GALENDEZ, G. C., RAEKE, L. H., SILVERSTEIN, N. J., SMOLLER, J. W., HILL, M. & GOFF, D. C. 2013. Randomized multicenter investigation of folate plus vitamin B12 supplementation in schizophrenia. JAMA psychiatry, 70, 481-489.
157. ROSE, J. E. & BEHM, F. M. 2013. Adapting smoking cessation treatment according to initial response to precessation nicotine patch. American Journal of Psychiatry, 170, 860-867.
158. ROSS, R. G., HUNTER, S. K., MCCARTHY, L., BEULER, J., HUTCHISON, A. K., WAGNER, B. D., LEONARD, S., STEVENS, K. E. & FREEDMAN, R. 2013. Perinatal choline effects on neonatal pathophysiology related to later schizophrenia risk. American Journal of Psychiatry, 170, 290-298.
159. ROTHBAUM, B. O., PRICE, M., JOVANOVIC, T., NORRHOLM, S. D., GERARDI, M., DUNLOP, B., DAVIS, M., BRADLEY, B., DUNCAN, E. J. & RIZZO, A. 2014. A Randomized, double-blind evaluation of d-cycloserine or alprazolam combined with virtual reality exposure therapy for posttraumatic stress disorder in Iraq and Afghanistan war veterans. American Journal of Psychiatry, 171, 640-648.
160. RUSH, A. J., TRIVEDI, M. H., STEWART, J. W., NIERENBERG, A. A., FAVA, M., KURIAN, B. T., WARDEN, D., MORRIS, D. W., LUTHER, J. F. & HUSAIN, M. M. 2011. Combining Medications to Enhance Depression Outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. American Journal of Psychiatry, 168, 689-701.
161. SACKTOR, N., MIYAHARA, S., DENG, L., EVANS, S., SCHIFITTO, G., COHEN, B., PAUL, R., ROBERTSON, K., JAROCKI, B. & SCARSI, K. 2011. Minocycline treatment for HIV-associated cognitive impairment Results from a randomized trial. Neurology, 77, 1135-1142.
162. SALLOWAY, S., SPERLING, R., FOX, N. C., BLENNOW, K., KLUNK, W., RASKIND, M., SABBAGH, M., HONIG, L. S., PORSTEINSSON, A. P. & FERRIS, S. 2014. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. New England Journal of Medicine, 370, 322-333.
163. SANDERCOCK, P., WARDLAW, J. M., LINDLEY, R. I., DENNIS, M., COHEN, G., MURRAY, G., INNES, K., VENABLES, G., CZLONKOWSKA, A. & KOBAYASHI, A. 2012. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. Lancet, 379, 2352-2363.
164. SANO, M., BELL, K., GALASKO, D., GALVIN, J., THOMAS, R., VAN DYCK, C. & AISEN, P. 2011. A randomized, double-blind, placebo-controlled trial of simvastatin to treat Alzheimer disease. Neurology, 77, 556-563.
165. SAVER, J. L., JAHAN, R., LEVY, E. I., JOVIN, T. G., BAXTER, B., NOGUEIRA, R. G., CLARK, W., BUDZIK, R. & ZAIDAT, O. O. 2012. Solitaire flow restoration device versus the Merci Retriever in patients with acute ischaemic stroke (SWIFT): a randomised, parallel-group, non-inferiority trial. The Lancet, 380, 1241-1249.
166. SCHAPIRA, A., BARONE, P., HAUSER, R., MIZUNO, Y., RASCOL, O., BUSSE, M., SALIN, L., JUHEL, N. & POEWE, W. 2011. Extended-release pramipexole in advanced Parkinson disease A randomized controlled trial. Neurology, 77, 767-774.
167. SCHAPIRA, A. H., MCDERMOTT, M. P., BARONE, P., COMELLA, C. L., ALBRECHT, S., HSU, H. H., MASSEY, D. H., MIZUNO, Y., POEWE, W. & RASCOL, O. 2013. Pramipexole in patients

with early Parkinson's disease (PROUD): a randomised delayed-start trial. The Lancet Neurology, 12, 747-755.
168. SCHNEIER, F. R., NERIA, Y., PAVLICOVA, M., HEMBREE, E., SUH, E. J., AMSEL, L. & MARSHALL, R. D. 2012. Combined prolonged exposure therapy and paroxetine for PTSD related to the World Trade Center attack: a randomized controlled trial. American Journal of Psychiatry, 169, 80-88.
169. SCHUEPBACH, W., RAU, J., KNUDSEN, K., VOLKMANN, J., KRACK, P., TIMMERMANN, L., HÄLBIG, T., HESEKAMP, H., NAVARRO, S. & MEIER, N. 2013. Neurostimulation for Parkinson's disease with early motor complications. New England Journal of Medicine, 368, 610-622.
170. SCOTT, P. A., MEURER, W. J., FREDERIKSEN, S. M., KALBFLEISCH, J. D., XU, Z., HAAN, M. N., SILBERGLEIT, R. & MORGENSTERN, L. B. 2012. A multilevel intervention to increase community hospital use of alteplase for acute stroke (INSTINCT): a cluster-randomised controlled trial. The Lancet Neurology.
171. SHALEV, A. Y., ANKRI, Y., ISRAELI-SHALEV, Y., PELEG, T., ADESSKY, R. & FREEDMAN, S. 2012. Prevention of posttraumatic stress disorder by early treatment: results from the Jerusalem Trauma Outreach And Prevention study. Archives of general psychiatry, 69, 166-176.
172. SHARPE, M., WALKER, J., WILLIAMS, C., STONE, J., CAVANAGH, J., MURRAY, G., BUTCHER, I., DUNCAN, R., SMITH, S. & CARSON, A. 2011. Guided self-help for functional (psychogenic) symptoms A randomized controlled efficacy trial. Neurology, 77, 564-572.
173. SIGMON, S. C., DUNN, K. E., SAULSGIVER, K., PATRICK, M. E., BADGER, G. J., HEIL, S. H., BROOKLYN, J. R. & HIGGINS, S. T. 2013. A randomized, double-blind evaluation of buprenorphine taper duration in primary prescription opioid abusers. JAMA psychiatry, 70, 1347-1354.
174. SILBERGLEIT, R., DURKALSKI, V., LOWENSTEIN, D., CONWIT, R., PANCIOLI, A., PALESCH, Y. & BARSAN, W. 2012. Intramuscular versus intravenous therapy for prehospital status epilepticus. New England Journal of Medicine, 366, 591-600.
175. SILBERSTEIN, S., DODICK, D., LINDBLAD, A., HOLROYD, K., HARRINGTON, M., MATHEW, N. & HIRTZ, D. 2012. Randomized, placebo-controlled trial of propranolol added to topiramate in chronic migraine. Neurology, 78, 976-984.
176. SILVER, B., ZAMAN, I. F., ASHRAF, K., MAJED, Y., NORWOOD, E. M., SCHUH, L. A., SMITH, B. J., SMITH, R. E. & SCHULTZ, L. R. 2012. A randomized trial of decision-making in asymptomatic carotid stenosis. Neurology, 78, 315-21.
177. SIMPSON, H. B., FOA, E. B., LIEBOWITZ, M. R., HUPPERT, J. D., CAHILL, S., MAHER, M. J., MCLEAN, C. P., BENDER, J., MARCUS, S. M. & WILLIAMS, M. T. 2013. Cognitive-behavioral therapy vs risperidone for augmenting serotonin reuptake inhibitors in obsessive-compulsive disorder: a randomized clinical trial. JAMA psychiatry, 70, 1190-1199.
178. SMYKE, A. T., ZEANAH, C. H., GLEASON, M. M., DRURY, S. S., FOX, N. A., NELSON, C. A. & GUTHRIE, D. 2012. A randomized controlled trial comparing foster care and institutional care for children with signs of reactive attachment disorder. American Journal of Psychiatry, 169, 508-514.
179. SOMOZA, E. C., WINSHIP, D., GORODETZKY, C. W., LEWIS, D., CIRAULO, D. A., GALLOWAY, G. P., SEGAL, S. D., SHEEHAN, M., ROACHE, J. D., BICKEL, W. K., JASINSKI, D., WATSON, D. W., MILLER, S. R., SOMOZA, P. & WINHUSEN, T. 2013. A multisite, double-blind, placebo-controlled clinical trial to evaluate the safety and efficacy of vigabatrin for treating cocaine dependence. JAMA Psychiatry, 70, 630-7.
180. SORENSEN, P. S., LYCKE, J., ERÄLINNA, J.-P., EDLAND, A., WU, X., FREDERIKSEN, J. L., OTURAI, A., MALMESTRÖM, C., STENAGER, E. & SELLEBJERG, F. 2011. Simvastatin as add-on therapy to interferon beta-1a for relapsing-remitting multiple sclerosis (SIMCOMBIN study): a placebo-controlled randomised phase 4 trial. The Lancet Neurology, 10, 691-701.
181. SORMANI, M., LI, D., BRUZZI, P., STUBINSKI, B., CORNELISSE, P., ROCAK, S. & DE STEFANO, N. 2011. Combined MRI lesions and relapses as a surrogate for disability in multiple sclerosis. Neurology, 77, 1684-1690.

182. SOROND, F. A., HURWITZ, S., SALAT, D. H., GREVE, D. N. & FISHER, N. D. 2013. Neurovascular coupling, cerebral white matter integrity, and response to cocoa in older people. Neurology, 81, 904-909.
183. STANGIER, U., HILLING, C., HEIDENREICH, T., RISCH, A. K., BAROCKA, A., SCHLÖSSER, R., KRONFELD, K., RUCKES, C., BERGER, H. & RÖSCHKE, J. 2013. Maintenance Cognitive-Behavioral Therapy and Manualized Psychoeducation in the Treatment of Recurrent Depression: A Multicenter Prospective Randomized Controlled Trial. American Journal of Psychiatry, 170, 624-632.
184. STANGIER, U., SCHRAMM, E., HEIDENREICH, T., BERGER, M. & CLARK, D. M. 2011. Cognitive therapy vs interpersonal psychotherapy in social anxiety disorder: a randomized controlled trial. Archives of general psychiatry, 68, 692.
185. STROUP, T. S., MCEVOY, J. P., RING, K. D., HAMER, R. H., LAVANGE, L. M., SWARTZ, M. S., ROSENHECK, R. A., PERKINS, D. O., NUSSBAUM, A. M. & LIEBERMAN, J. A. 2011. A randomized trial examining the effectiveness of switching from olanzapine, quetiapine, or risperidone to aripiprazole to reduce metabolic risk: comparison of antipsychotics for metabolic problems (CAMP). American Journal of Psychiatry, 168, 947-956.
186. STURKENBOOM, I. H., GRAFF, M. J., HENDRIKS, J., VEENHUIZEN, Y., MUNNEKE, M., BLOEM, B. R. & DER SANDEN, M. W. 2014. Efficacy of occupational therapy for patients with Parkinson's disease: a randomised controlled trial. The Lancet Neurology, 13, 557-566.
187. TARIOT, P. N., SCHNEIDER, L. S., CUMMINGS, J., THOMAS, R. G., RAMAN, R., JAKIMOVICH, L. J., LOY, R., BARTOCCI, B., FLEISHER, A. & ISMAIL, M. S. 2011. Chronic divalproex sodium to attenuate agitation and clinical progression of Alzheimer disease. Archives of General Psychiatry, 68, 853.
188. THORNICROFT, G., FARRELLY, S., SZMUKLER, G., BIRCHWOOD, M., WAHEED, W., FLACH, C., BARRETT, B., BYFORD, S., HENDERSON, C. & SUTHERBY, K. 2013. Clinical outcomes of joint crisis plans to reduce compulsory treatment for people with psychosis: a randomised controlled trial. The Lancet, 381, 1634-1641.
189. TORRENT, C., DEL MAR BONNIN, C., MARTÍNEZ-ARÁN, A., VALLE, J., AMANN, B. L., GONZÁLEZ-PINTO, A., CRESPO, J. M., IBÁÑEZ, Á., GARCIA-PORTILLA, M. P. & TABARÉS-SEISDEDOS, R. 2013. Efficacy of functional remediation in bipolar disorder: a multicenter randomized controlled study. American Journal of Psychiatry, 170, 852-859.
190. TRENKWALDER, C., BENEŠ, H., GROTE, L., GARCÍA-BORREGUERO, D., HÖGL, B., HOPP, M., BOSSE, B., OKSCHE, A., REIMER, K. & WINKELMANN, J. 2013. Prolonged release oxycodone–naloxone for treatment of severe restless legs syndrome after failure of previous treatment: a double-blind, randomised, placebo-controlled trial with an open-label extension. The Lancet Neurology, 12, 1141-1150.
191. TYRER, P., COOPER, S., SALKOVSKIS, P., TYRER, H., CRAWFORD, M., BYFORD, S., DUPONT, S., FINNIS, S., GREEN, J. & MCLAREN, E. 2014. Clinical and cost-effectiveness of cognitive behaviour therapy for health anxiety in medical patients: a multicentre randomised controlled trial. The Lancet, 383, 219-225.
192. UNDERWOOD, M., LAMB, S. E., ELDRIDGE, S., SHEEHAN, B., SLOWTHER, A.-M., SPENCER, A., THOROGOOD, M., ATHERTON, N., BREMNER, S. A. & DEVINE, A. 2013. Exercise for depression in elderly residents of care homes: a cluster-randomised controlled trial. The Lancet, 382, 41-49.
193. VAN DER VAART, T., PLASSCHAERT, E., RIETMAN, A. B., RENARD, M., OOSTENBRINK, R., VOGELS, A., DE WIT, M.-C. Y., DESCHEEMAEKER, M.-J., VERGOUWE, Y. & CATSMAN-BERREVOETS, C. E. 2013. Simvastatin for cognitive deficits and behavioural problems in patients with neurofibromatosis type 1 (NF1-SIMCODA): a randomised, placebo-controlled trial. The Lancet Neurology, 12, 1076-1083.
194. VELLAS, B., COLEY, N., OUSSET, P.-J., BERRUT, G., DARTIGUES, J.-F., DUBOIS, B., GRANDJEAN, H., PASQUIER, F., PIETTE, F. & ROBERT, P. 2012. Long-term use of standardised ginkgo biloba extract for the prevention of Alzheimer's disease (GuidAge): a randomised placebo-controlled trial. The Lancet Neurology.

195. VITIELLO, B., ELLIOTT, G. R., SWANSON, J. M., ARNOLD, L. E., HECHTMAN, L., ABIKOFF, H., MOLINA, B. S., WELLS, K., WIGAL, T. & JENSEN, P. S. 2012. Blood pressure and heart rate over 10 years in the multimodal treatment study of children with ADHD. American Journal of Psychiatry, 169, 167-177.
196. VOLKMANN, J., WOLTERS, A., KUPSCH, A., MÜLLER, J., KÜHN, A. A., SCHNEIDER, G.-H., POEWE, W., HERING, S., EISNER, W. & MÜLLER, J.-U. 2012. Pallidal deep brain stimulation in patients with primary generalised or segmental dystonia: 5-year follow-up of a randomised trial. The Lancet Neurology.
197. WANG, Y., WANG, Y., ZHAO, X., LIU, L., WANG, D., WANG, C., WANG, C., LI, H., MENG, X. & CUI, L. 2013. Clopidogrel with aspirin in acute minor stroke or transient ischemic attack. New England Journal of Medicine, 369, 11-19.
198. WEAVER, F. M., FOLLETT, K. A., STERN, M., LUO, P., HARRIS, C. L., HUR, K., MARKS, W. J., ROTHLIND, J., SAGHER, O. & MOY, C. 2012. Randomized trial of deep brain stimulation for Parkinson disease Thirty-six-month outcomes. Neurology, 79, 55-65.
199. WEISS, R. D., POTTER, J. S., FIELLIN, D. A., BYRNE, M., CONNERY, H. S., DICKINSON, W., GARDIN, J., GRIFFIN, M. L., GOUREVITCH, M. N. & HALLER, D. L. 2011. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence: a 2-phase randomized controlled trial. Archives of General Psychiatry, archgenpsychiatry. 2011.121 v1.
200. WEISZ, J. R., CHORPITA, B. F., PALINKAS, L. A., SCHOENWALD, S. K., MIRANDA, J., BEARMAN, S. K., DALEIDEN, E. L., UGUETO, A. M., HO, A. & MARTIN, J. 2012. Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. Archives of general psychiatry, 69, 274-282.
201. WEST, R., ZATONSKI, W., CEDZYNSKA, M., LEWANDOWSKA, D., PAZIK, J., AVEYARD, P. & STAPLETON, J. 2011. Placebo-controlled trial of cytisine for smoking cessation. New England Journal of Medicine, 365, 1193-1200.
202. WETHERELL, J. L., PETKUS, A. J., WHITE, K. S., NGUYEN, H., KORNBLITH, S., ANDREESCU, C., ZISOOK, S. & LENZE, E. J. 2013. Antidepressant medication augmented with cognitive-behavioral therapy for generalized anxiety disorder in older adults. American Journal of Psychiatry, 170, 782-789.
203. WILES, N., THOMAS, L., ABEL, A., RIDGWAY, N., TURNER, N., CAMPBELL, J., GARLAND, A., HOLLINGHURST, S., JERROM, B. & KESSLER, D. 2013. Cognitive behavioural therapy as an adjunct to pharmacotherapy for primary care based patients with treatment resistant depression: results of the CoBalT randomised controlled trial. The Lancet, 381, 375-384.
204. WILHELM, S., PETERSON, A. L., PIACENTINI, J., WOODS, D. W., DECKERSBACH, T., SUKHODOLSKY, D. G., CHANG, S., LIU, H., DZIURA, J. & WALKUP, J. T. 2012. Randomized trial of behavior therapy for adults with Tourette syndrome. Archives of general psychiatry, 69, 795-803.
205. WU, R.-R., JIN, H., GAO, K., TWAMLEY, E. W., OU, J.-J., SHAO, P., WANG, J., GUO, X.-F., DAVIS, J. M. & CHAN, P. K. 2012. Metformin for treatment of antipsychotic-induced amenorrhea and weight gain in women with first-episode schizophrenia: a double-blind, randomized, placebo-controlled study. American Journal of Psychiatry, 169, 813-821.
206. ZAJICEK, J., BALL, S., WRIGHT, D., VICKERY, J., NUNN, A., MILLER, D., CANO, M. G., MCMANUS, D., MALLIK, S. & HOBART, J. 2013. Effect of dronabinol on progression in progressive multiple sclerosis (CUPID): a randomised, placebo-controlled trial. The Lancet Neurology, 12, 857-865.
207. ZIMMERMANN, R., GSCHWANDTNER, U., BENZ, N., HATZ, F., SCHINDLER, C., TAUB, E. & FUHR, P. 2014. Cognitive training in Parkinson disease Cognition-specific vs nonspecific computer training. Neurology, 82, 1219-1226.
208. ZINKSTOK, S. M. & ROOS, Y. B. 2012. Early administration of aspirin in patients treated with alteplase for acute ischaemic stroke: a randomised controlled trial. The Lancet, 380, 731-737.

209. ZIPFEL, S., WILD, B., GROß, G., FRIEDERICH, H.-C., TEUFEL, M., SCHELLBERG, D., GIEL, K. E., DE ZWAAN, M., DINKEL, A. & HERPERTZ, S. 2014. Focal psychodynamic therapy, cognitive behaviour therapy, and optimised treatment as usual in outpatients with anorexia nervosa (ANTOP study): randomised controlled trial. The Lancet, 383, 127-137.

Appendix 3

Published manuscript of methods to adjust for multiple comparisons

Vickerstaff *et al. BMC Medical Research Methodology* (2019) 19:129
https://doi.org/10.1186/s12874-019-0754-4

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes

Victoria Vickerstaff[1,2*], Rumana Z. Omar[2] and Gareth Ambler[2]

Abstract

Background: Multiple primary outcomes may be specified in randomised controlled trials (RCTs). When analysing multiple outcomes it's important to control the family wise error rate (FWER). A popular approach to do this is to adjust the *p*-values corresponding to each statistical test used to investigate the intervention effects by using the Bonferroni correction. It's also important to consider the power of the trial to detect true intervention effects. In the context of multiple outcomes, depending on the clinical objective, the power can be defined as: *'disjunctive power'*, the probability of detecting at least one true intervention effect across all the outcomes or *'marginal power'* the probability of finding a true intervention effect on a nominated outcome.
We provide practical recommendations on which method may be used to adjust for multiple comparisons in the sample size calculation and the analysis of RCTs with multiple primary outcomes. We also discuss the implications on the sample size for obtaining 90% disjunctive power and 90% marginal power.

Methods: We use simulation studies to investigate the disjunctive power, marginal power and FWER obtained after applying Bonferroni, Holm, Hochberg, Dubey/Armitage-Parmar and Stepdown-minP adjustment methods. Different simulation scenarios were constructed by varying the number of outcomes, degree of correlation between the outcomes, intervention effect sizes and proportion of missing data.

Results: The Bonferroni and Holm methods provide the same disjunctive power. The Hochberg and Hommel methods provide power gains for the analysis, albeit small, in comparison to the Bonferroni method. The Stepdown-minP procedure performs well for complete data. However, it removes participants with missing values prior to the analysis resulting in a loss of power when there are missing data. The sample size requirement to achieve the desired disjunctive power may be smaller than that required to achieve the desired marginal power. The choice between whether to specify a disjunctive or marginal power should depend on the clincial objective.

Keywords: Multiple comparison methods, Multiple outcome, Sample size, Statistical analysis, Randomised controlled trials

* Correspondence: v.vickerstaff@ucl.ac.uk
[1]Marie Curie Palliative Care Research Department, Division of Psychiatry, University College London, Gower Street, London WC1E 6BT, UK
[2]Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

Background

Multiple primary outcomes may be specified in a randomised controlled trial (RCT) when it is not possible to use a single outcome to fully characterise the effect of an intervention on a disease process [1–3]. The use of multiple primary outcomes (or 'endpoints') is becoming increasingly common in RCTs. For example, a third of neurology and psychiatry trials use multiple primary outcomes [4]. Data on two primary outcomes (abstinence and time to dropout from treatment) were collected in a trial evaluating the effectiveness of a behavioural intervention for substance abuse [5] and data on four primary outcomes were collected in a trial evaluating a multidisciplinary intervention in patients following a stroke [6]. Typically, these outcomes are correlated and often one or more of the outcomes has missing values.

Typically multiple statistical tests are performed to investigate the effectiveness of the intervention on each outcome. If two outcomes are analysed independently of each other at the nominal significance level of 0.05, then the probability of finding at least one false positive significant results increases to 0.098. This probability is known as the familywise error rate, 'FWER'. One approach to control the FWER to its desired level is to adjust the p-values corresponding to each statistical test used to investigate the intervention effects. Many adjustments have been proposed including the Bonferroni [7], Holm [8], Hochberg [9], Hommel [10] and Dubey/Armitage-Parmar [11] methods. Once the p-values have been adjusted, they can be compared to the nominal significance level. For example in the trial on substance abuse [5], two unadjusted p-values: 0.010,0.002 were reported. If the Bonferroni method was used, the p-values could have been adjusted to 0.020, 0.004 and compared to the significance level α of 0.05. Alternatively, the significance level could be adjusted (to $0.05/2 = 0.025$ in this example) and compared to the unadjusted p-values.

In clinical trials, it is also important to consider the power of the tests to detect an intervention effect. In the context of multiple outcomes, the power of the study can be defined in a number of ways depending on the clinical objective of the trial: i) 'disjunctive power', ii) 'conjunctive power' or iii) 'marginal power' [12].

The disjunctive power (or minimal power [13]) is the probability of finding at least one true intervention effect across all of the outcomes [12, 14]. The conjunctive power (or maximal power [13]) is the probability of finding a true intervention effect on all outcomes [14]. It may be noted that the disjunctive and conjunctive power have previously been referred to as 'multiple' and 'complete' power respectively [13]. The marginal (or individual) power is the probability of finding a true intervention effect on a particular outcome and is calculated separately for each outcome. When the clinical objective is to detect an intervention effect for at least one of the outcomes the disjunctive power and marginal power are recommended whereas the conjunctive power is recommended when the clinical objective is to detect an intervention effect on all the outcomes [12, 14]. In this paper, we are focusing on the former clinical objective and therefore we focus on disjunctive and marginal power.

The power requirements of a trial should match the clinical objective which needs to be pre-specified when designing the study and the sample size calculation should be performed accordingly. In current practice, the sample size calculations for trials often focus on the marginal power for each outcome. An approach that has been recommended and is often used in trials is to calculate the sample size separately for each of the primary outcomes by applying a Bonferroni correction to adjust the significance level [15]. The largest value of the sample size is then considered as the final sample size for the trial [16].

Missing outcome data are common in RCTs [17] which will inevitably reduce the power and efficiency of the study [18] which may result in failure to detect true intervention effects as statistically significant.

When using multiple primary outcomes, there is limited guidance as to which method(s) should be used to take account of multiplicity in the sample size calculation and during the statistical analysis.

Some studies have compared a selection of methods which adjust p-values to account for multiplicity to handle multiple outcomes in trials. Sankoh, Huque and Dubey [11] compare a selection of adjustment methods for statistical analysis in terms of FWER but they do not evaluate the methods with respect to the power obtained. Blakesley et al. discuss both FWER and power requirements for selected methods for a large number of outcomes with varying degrees of correlation [19]. Lafaye de Micheaux provide formulae to calculate the power and sample size for multiple outcomes [20] which require several assumptions to be made about the outcomes, including normality and whether the covariance matrix between the outcomes is known or not. They discuss global testing procedures, including the Hotelling T^2 method. None of these studies have investigated the adjustment methods in the presence of missing data.

There is limited literature discussing the sample size requirements for clinical trials with multiple primary outcomes where the clinical objective is to detect an intervention effect for at least one of the outcomes.

Dmitrienko, Tamhane and Bretz [14] and Senn and Bretz [13] provide some discussion regarding the sample size in the context of multiple outcomes. However, neither discuss sample size in the context of which adjustment method should be used and they do not provide a comparative table depending on the type of desired power to show implications on the required sample sizes.

In this paper, we compare easy to use methods to adjust p-values in terms of FWER and power, when investigating two, three and four outcomes in presence of complete outcome data and outcome data with missing values. We also consider a range of correlations between the outcomes. We consider both marginal and disjunctive power. Based on our findings, we provide practical recommendations on the adjustment methods which could be used for the sample size calculation and analysis of RCTs with multiple primary outcome. We also present tables showing the implications of using the marginal and disjunctive power on the required sample size for a trial under different scenarios.

Methods

We assume that we have a two-arm trial in which there are M primary outcomes. We are interested in testing the null hypotheses H_j $(j = 1, \ldots, M)$ that there is no intervention effect on the nominated outcomes. The test statistics t_j are used to test the null hypotheses H_j. Further suppose that there is an overall null hypothesis $H(M) = \cap_{j=1}^{M} H_j$. Under this overall hypothesis, the joint test statistic $(t_1, \ldots, t_M)$ has a M-variate distribution. We denote p_j as the marginal, unadjusted p-values obtained from the appropriate statistical test associated with analysing each outcome separately in a univariate framework. For example, when analysing continuous outcomes, an unpaired Student's t-test may be used or when analysing binary outcomes a Chi-squared test may be used to investigate the intervention. To control the FWER a correction method is then applied to the unadjusted p-values (p_j). We compare the following commonly used adjustment methods in this paper: Šidák, Bonferroni, Holm, Hochberg and Hommel. In addition, we consider the Dubey/Armitage-Parmar (D/AP) adjustment and Stepdown minP resampling procedure which take account of the pairwise correlation between the outcomes.

The method proposed by Šidák is defined as $\overset{\breve{}}{p}_j^{\mathrm{Si}} = 1-(1-p_j)^M$. Equivalently, the significance level could be adjusted to $\overset{\breve{}}{\alpha}^{\mathrm{Si}} = 1-(1-\alpha)^{1/M}$, where α is the unadjusted significance level. Under the assumption that the outcomes are independent, the adjustment can be derived as

$$P(\textit{no Type I error on } \mathbf{1} \textit{ test}) = 1-\overset{\breve{}}{\alpha}^{\mathrm{Si}},$$

$$\rightarrow P(\textit{no Type I error on } \mathbf{M} \textit{ tests}) = (1-\overset{\breve{}}{\alpha}^{\mathrm{Si}})^{M},$$

$$\rightarrow P(\textit{atleast one Type I error on } \mathrm{M} \textit{ tests}) = 1-(1-\overset{\breve{}}{\alpha}^{\mathrm{Si}})^{M} = \alpha.$$

The Bonferroni method is the most common approach to account for multiplicity due to its simplicity. In this method, the unadjusted p-values p_j are multiplied by the number of primary outcome. The Dubey/Armitage-Parmar (D/AP) is an ad-hoc method based on the Šidák method, which takes into account the correlation between the outcomes [11]. The adjusted p-value is $p_j^{adj} = 1-(1-p_j)^{g(j)}$ where $g(j) = M^{1-mean\ \rho(j)}$ and $mean\ \rho(j)$ is the mean correlation between the j^{th} outcome and the remaining $M-1$ outcomes. When using this method in the analysis of multiple outcomes, the mean correlation may be estimated from the data. There has been little theoretical work to assess the performance of this approach [11].One of the nice properties of the D/AP procedure, which may have contributed to its development, is that when the average of the correlation coefficients is zero, the D/AP adjustment is according to the Bonferroni test, and when the average correlation coefficient is one, the D/AP adjusted and the unadjusted p-values are the same. The Holm method [8] involves a step-down method, whereby the unadjusted p-values are ordered from smallest $p_{(1)}$ to largest $p_{(M)}$ and each unadjusted p-value is adjusted as $p_{(k)}^{Holm} = (M-k+1)\, p_{(k)}$, where $k = 1, \ldots M$ is the rank of the corresponding p-value. Then starting with the most significant p-value (smallest p-value), each adjusted p-value is compared to the nominal significance level, until a p-value *greater* than the significance level is observed after which the method stops [21]. The Hochberg step-up method [9] is similar to the Holm step-down method but works in the other direction. For this method, the unadjusted p-values are ranked from largest $p_{(1)}$ to smallest $p_{(M)}$ and adjusted as $p_{(k)}^{Hoch} = (M-k+1)\, p_{(k)}$. Starting with the least significant p-value (largest p-value), each adjusted p-value is compared to the pre-specified significance level, until a p-value *lower* than the significance level is observed after which the method stops [21]. Contrary to the Šidák based approaches, this is a semiparametric method meaning the FWER is only controlled when the joint distribution of the hypotheses test statistics is known, most commonly multivariate normal [22]. The Hommel method [10] is another data-driven stepwise method.

Dmitrienko, Tamhane and Bretz [14] and Senn and Bretz [13] provide some discussion regarding the sample size in the context of multiple outcomes. However, neither discuss sample size in the context of which adjustment method should be used and they do not provide a comparative table depending on the type of desired power to show implications on the required sample sizes.

In this paper, we compare easy to use methods to adjust p-values in terms of FWER and power, when investigating two, three and four outcomes in presence of complete outcome data and outcome data with missing values. We also consider a range of correlations between the outcomes. We consider both marginal and disjunctive power. Based on our findings, we provide practical recommendations on the adjustment methods which could be used for the sample size calculation and analysis of RCTs with multiple primary outcome. We also present tables showing the implications of using the marginal and disjunctive power on the required sample size for a trial under different scenarios.

Methods

We assume that we have a two-arm trial in which there are M primary outcomes. We are interested in testing the null hypotheses H_j $(j = 1, \ldots, M)$ that there is no intervention effect on the nominated outcomes. The test statistics t_j are used to test the null hypotheses H_j. Further suppose that there is an overall null hypothesis $H(M) = \cap_{j=1}^{M} H_j$. Under this overall hypothesis, the joint test statistic $(t_1, \ldots, t_M)$ has a M-variate distribution. We denote p_j as the marginal, unadjusted p-values obtained from the appropriate statistical test associated with analysing each outcome separately in a univariate framework. For example, when analysing continuous outcomes, an unpaired Student's t-test may be used or when analysing binary outcomes a Chi-squared test may be used to investigate the intervention. To control the FWER a correction method is then applied to the unadjusted p-values (p_j). We compare the following commonly used adjustment methods in this paper: Šidák, Bonferroni, Holm, Hochberg and Hommel. In addition, we consider the Dubey/Armitage-Parmar (D/AP) adjustment and Stepdown minP resampling procedure which take account of the pairwise correlation between the outcomes.

The method proposed by Šidák is defined as $\overset{\vee}{p}_j^{\mathrm{Si}} = 1-(1-p_j)^M$. Equivalently, the significance level could be adjusted to $\overset{\vee}{\alpha}^{\mathrm{Si}} = 1-(1-\alpha)^{1/M}$, where α is the unadjusted significance level. Under the assumption that the outcomes are independent, the adjustment can be derived as

$$P(\textit{no Type I error on } \mathbf{1} \textit{ test}) = 1-\overset{\vee}{\alpha}^{\mathrm{Si}},$$

$$\rightarrow P(\textit{no Type I error on } \mathbf{M} \textit{ tests}) = (1-\overset{\vee}{\alpha}^{\mathrm{Si}})^{M},$$

$$\rightarrow P(\textit{at least one Type I error on M tests}) = 1-(1-\overset{\vee}{\alpha}^{\mathrm{Si}})^{M} = \alpha.$$

The Bonferroni method is the most common approach to account for multiplicity due to its simplicity. In this method, the unadjusted p-values p_j are multiplied by the number of primary outcome. The Dubey/Armitage-Parmar (D/AP) is an ad-hoc method based on the Šidák method, which takes into account the correlation between the outcomes [11]. The adjusted p-value is $p_j^{adj} = 1-(1-p_j)^{g(j)}$ where $g(j) = M^{1-mean\ \rho(j)}$ and $mean\ \rho(j)$ is the mean correlation between the j^{th} outcome and the remaining $M-1$ outcomes. When using this method in the analysis of multiple outcomes, the mean correlation may be estimated from the data. There has been little theoretical work to assess the performance of this approach [11].One of the nice properties of the D/AP procedure, which may have contributed to its development, is that when the average of the correlation coefficients is zero, the D/AP adjustment is according to the Bonferroni test, and when the average correlation coefficient is one, the D/AP adjusted and the unadjusted p-values are the same. The Holm method [8] involves a step-down method, whereby the unadjusted p-values are ordered from smallest $p_{(1)}$ to largest $p_{(M)}$ and each unadjusted p-value is adjusted as $p_{(k)}^{Holm} = (M-k+1)\, p_{(k)}$, where $k = 1, \ldots M$ is the rank of the corresponding p-value. Then starting with the most significant p-value (smallest p-value), each adjusted p-value is compared to the nominal significance level, until a p-value *greater* than the significance level is observed after which the method stops [21]. The Hochberg step-up method [9] is similar to the Holm step-down method but works in the other direction. For this method, the unadjusted p-values are ranked from largest $p_{(1)}$ to smallest $p_{(M)}$ and adjusted as $p_{(k)}^{Hoch} = (M-k+1)\, p_{(k)}$. Starting with the least significant p-value (largest p-value), each adjusted p-value is compared to the pre-specified significance level, until a p-value *lower* than the significance level is observed after which the method stops [21]. Contrary to the Šidák based approaches, this is a semiparametric method meaning the FWER is only controlled when the joint distribution of the hypotheses test statistics is known, most commonly multivariate normal [22]. The Hommel method [10] is another data-driven stepwise method.

For this method, the unadjusted p-values are ranked from largest $p_{(M)}$ to smallest $p_{(1)}$. Then let l be the largest integer for which $p_{(M-l+j)} > \frac{j\alpha}{l}$ or all $j = 1, \ldots l$. If no such j exists then all outcomes can be deemed statistically significant; otherwise, all outcomes with $p_i \leq \frac{\alpha}{j}$ may be deemed statistically significant, where $j = 1, \ldots, M$; $i = 1, \ldots, M$. To control the FWER, the Hommel method requires that the joint distribution of the overall hypothesis test statistic is known.

Another step-down method to adjust p-values is the 'Stepdown minP' procedure [23, 24]. Unlike the previous methods, it does not make any assumptions regarding the distribution of the joint test statistic. Instead it attempts to approximate the true joint distribution by using a resampling approach. This method takes into account the correlation structure between the outcomes and therefore may yield more powerful tests compared to the other adjustment methods [25]. The Stepdown minP adjusted p-values are calculated as follows: 1) calculate the observed test statistics using the observed data set; 2) resample the data with replacement within each intervention group to obtain bootstrap resamples, compute the resampled test statistics for each resampled data set and construct the reference distribution using the centred and/or scaled resampled test statistics; 3) calculate the critical value of a level α test based on the upper α percentile of the reference distribution, or obtain the raw p-values by computing the proportion of bootstrapped test statistics that are as extreme or more extreme than the observed test statistic [26]. That is, the Stepdown minP adjusted p-value for the j^{th} outcome is defined as [24, 26] $p_j^{minP} = \max_{k=1,\ldots,j}\{\Pr((\min_{l=k,\ldots,M} p_l \leq p_k \mid H(M))\}$, where p_k is the unadjusted p-value for the k^{th} outcome, p_l is the unadjusted p-value for the l^{th} outcome ($l = k, \ldots, M$), and $H(M)$ is the overall null hypothesis.

Although, the resampling based methods have previously been recommended for clinical trials with multiple outcomes they are not widely used in practice [25]. The Stepdown minP has been shown to perform well when compared to other resampling procedures [26] and was therefore investigated in this paper.

We perform a simulation study to evaluate the validity of these methods to account for potentially correlated multiple primary outcomes in the analysis and sample size of RCTs. We focus on two, three and four outcomes as a review of trials with multiple primary outcomes in the psychiatry and neurology field found that the majority of the trials had considered two primary outcomes [4]. Additionally, it has been recommended that a trial should have no more than four primary outcomes [27]. We estimate the family wise error rate (FWER), the disjunctive power to detect at least one intervention effect and the marginal power to detect an intervention effect on a nominated outcome in a variety of scenarios.

Simulation study

We used the following model to simulate values for two continuous outcomes $\boldsymbol{Y_i} = (Y_{i,1}, Y_{i,2})$,

$$\boldsymbol{Y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_i + \boldsymbol{\epsilon_i} \quad (2)$$

where x_i indicates whether the participant i received intervention or control, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^T$ is vector of the intervention effects for each outcome, $\boldsymbol{\epsilon_i}$ are errors which are realisations of a multivariate normal distribution $\boldsymbol{\epsilon_i} = (\epsilon_{i,1}, \epsilon_{i,2})^T \sim N((\begin{smallmatrix}0\\0\end{smallmatrix}), \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix})$, and $\rho \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. The model was also extended to simulate three and four continuous outcomes. When simulating three and four outcomes we specified compound symmetry, meaning that the correlation between any pair of outcomes is the same. We explored both uniform intervention effect sizes and varying effect sizes across outcomes. For the uniform intervention effect sizes, we specified an effect size of 0.35 for all outcomes, that is $\boldsymbol{\beta}_1 = (0.35, 0.35)^T$, $\boldsymbol{\beta}_1 = (0.35, 0.35, 0.35)^T$ or $\boldsymbol{\beta}_1 = (0.35, 0.35, 0.35, 0.35)^T$ for two, three and four outcomes scenarios respectively. This represents a medium effect size, which reflects the anticipated effect size in many RCTs [28]. For the varying intervention effect sizes, we specified that $\boldsymbol{\beta}_1 = (0.2, 0.4)^T$, $\boldsymbol{\beta}_1 = (0.2, 0.3, 0.4)^T$ or $\boldsymbol{\beta}_1 = (0.1, 0.2, 0.3, 0.4)^T$ for two, three and four outcomes scenarios respectively. We also explored the effect of skewed data by transforming the outcome data with uniform intervention effect sizes to have a gamma distribution with shape parameter = 2 and a scale parameter = 2. The gamma distribution is often used to model healthcare costs in clinical trials [29, 30] and may also be appropriate for skewed clinical outcomes.

We set the sample size to 260 participants, with an equal number of participants assigned to each arm. This provides 80% marginal power to detect a clinically important effect size of 0.35 for each outcome, using an unpaired Student's t-test and the significance level is unadjusted at 0.05. We introduced missing data under the assumption that the data were missing completely at random (MCAR). When simulating two outcomes, 15 and 25% of the observations in outcome 1 and 2 are missing respectively, and on average approximately 4% of the observations would be missing for both outcomes. When simulating three outcomes, 15% of the observations are missing in one outcome and 25% of the observations are missing in the other two outcomes. When simulating four outcomes, 15% of the observations are missing in two outcomes and 25% of the observations are missing in the other two outcomes. This proportion of missingness in outcomes is often observed in RCTs [31–34].

We estimated the FWER and disjunctive power by specifying no intervention effect ($\beta_{1j} = 0$) and an intervention effect ($\beta_{1j} \neq 0$), respectively, and calculating the proportion of times an intervention effect was observed on at least one of the outcomes. The marginal power was similarly estimated but we calculated the proportion of times an intervention effect was observed on the nominated outcome. For each scenario we ran 10,000 simulations. The simulations were run using R version 3.4.2. The Stepdown minP procedure was implemented using the NPC package.

We calculated the sample size based on disjunctive power using the R package "mpe" [35] and we calculated the sample size based on the marginal power using the R package "samplesize" [36]. The statistical methodology used for the sample size calculation in these packages is described in the Additional file 1.

Results

The Bonferroni and Holm methods lead to the same FWER and disjunctive power when analysing multiple primary outcomes. This is because both methods adjust the smallest *p*-value in the same way. Similarly, the Hochberg and Hommel methods lead to same FWER and disjunctive power when two primary outcomes are analysed and differences between these methods arise when analysing three or more outcomes.

Family wise error rate, FWER

The FWER obtained when evaluating two, three and four outcomes are displayed in Figs. 1, 2 and 3 respectively. Following on from the explanation above, the Holm and Hommel methods are not displayed in Fig. 1 and the Holm method is not displayed in Fig. 2 or 3. The results for the varying intervention effect sizes and skewed data are presented in the Additional file 1.

When there is correlation between the outcomes ($\rho \geq 0.2$), the D/AP method does not control the FWER. All other adjustment methods control the FWER in all scenarios. The Stepdown minP performs well in terms of FWER. Unlike the other methods, it maintains the error rate at 0.05 even when the strength of the correlation between the outcomes increases. Differences between the Bonferroni, Hochberg and Hommel methods arise when there is moderate correlation between outcomes ($\rho \geq 0.4$). The Hommel provides the FWER which is closest to 0.05, whilst being controlled, followed by Hochberg and then Bonferroni. Very similar results were observed when the outcomes followed a skewed distribution, consequently these results are presented in the Additional file 1.

Disjunctive power

Figures 1, 2 and 3 show that the disjunctive power decreases as the correlation between the outcomes increases for all approaches. We do not consider the power obtained when using the D/AP approach due to its poor performance in controlling the FWER. When there is no missing data, the Stepdown minP and Hommel approaches provide the highest disjunctive power. For weak to moderate correlation ($\rho = 0.2$ *to* 0.6) the Hommel method has slightly more disjunctive power, but the Stepdown minP performs better when there is strong correlation ($\rho = 0.8$). The Stepdown minP procedure gives the lowest power in the presence of missing data. This could be attributed to the fact that it uses listwise deletion removing participants with at least one missing value prior to the analysis thus resulting in a loss of power when there is missing data. As expected the Bonferroni method gives slightly lower power compared to the other methods for complete data but considerably out performs the Stepdown minP method when there is missing data. Very similar results were observed when the outcomes followed a skewed distribution.

When the intervention effect sizes varied, the differences observed between the methods were less pronounced. When using four outcomes with varying effect sizes, very similar disjunctive power were observed to that of constant effect sizes. When using the Hommel adjustment, higher disjunctive power was observed compared to the Holm and Bonferroni methods albeit by a very minimal amount.

Marginal power

The marginal power obtained for each outcome when using the different adjustment methods are shown in Table 1. In terms of marginal power, the Hommel adjustment was the most powerful method, followed closely by the Hochberg method. When two independent outcomes were analysed, a power of 76.8% was observed after applying a Hommel correction. The power decreased to 76.8 and 75.2% when three and four outcomes were analysed, respectively, after applying a Hommel correction. As expected the Bonferroni method was the most conservative method, providing the least power. However, contrary to popular belief, the Bonferroni method maintains similar levels of power as the strength of correlation increases.

When analysing two outcomes the percentage of simulations in which an intervention effect was observed on neither outcome, one outcome or both outcomes are shown in Table 2. When using the Holm method, a statistically significant intervention effect was observed on both outcomes in 48–58% of the simulations. This reduced to 36–48% of the simulations when using the Bonferroni method. As expected, when using the Hochberg adjustment the same results were observed as when using the Hommel adjustment. Compared to Holm,

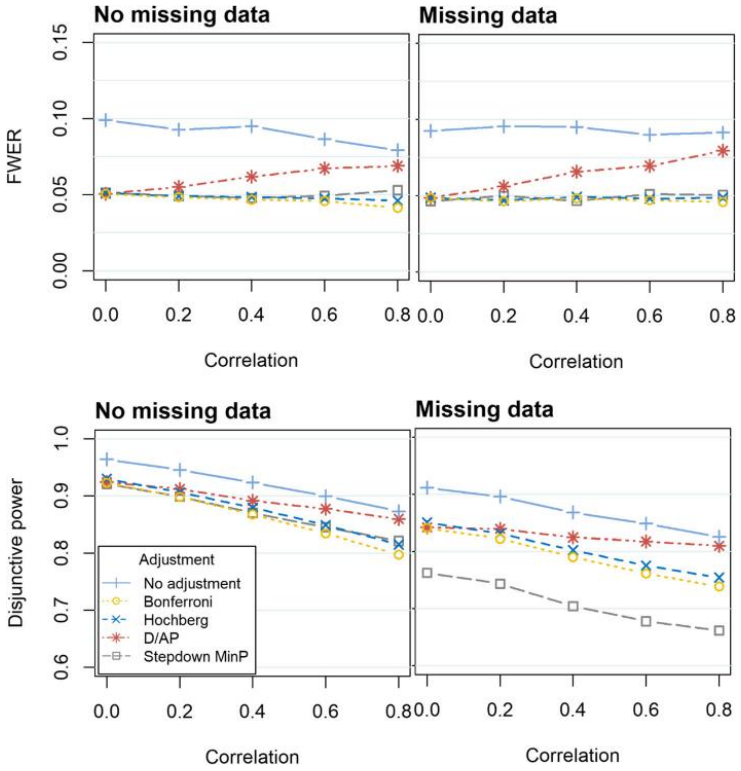


Fig. 1 The FWER (top) and disjunctive power (bottom) obtained when evaluating two continuous outcomes using a variety of methods to control the FWER. In the left hand graphs, there are no missing data. In the right hand graphs, the missing data are missing completely at random, with 15% missing in the first outcome and 25% missing in the second outcome ('Missing data'). The graphs display various degrees of correlation between the outcomes, ranging from $\rho = 0$ to $\rho = 0.8$. The Monte Carlo standard errors (MCSE) were similar across all methods. When there were no missing data, the MCSE was between 0.002–0.004 for the disjunctive power and 0.002–0.004 for the FWER. In the missing data scenario, the MCSE was between 0.002–0.003 for the disjunctive power and between 0.003–0.005 for the FWER.)

slightly higher percentages of simulations with two statistically significant intervention effects are observed when using Hochberg and Hommel.

Sample size calculation

We recommend the Bonferroni adjustment to be used for the sample size calculation when designing trials with multiple correlated outcomes since it can be applied easily by adjusting the significance level and it maintains the FWER to an acceptable level up to a correlation of 0.6 between outcomes. As the Hochberg and Hommel methods are data-driven, it is not clear how these more powerful approaches can be incorporated into the sample size calculation unless prior data are available. Determination of the required sample size using these methods may require simulation-based approach.

In Table 3, we present the required sample sizes to obtain 90% disjunctive power for trials with two outcomes for varying degrees of correlations between the outcomes ($\rho = \{0.2, 0.4, 0.6, 0.8\}$). For these calculations, we specified that there is equal allocation of participants between the intervention arms. To calculate the sample size a priori information on the degree of correlation between the outcomes is required. More details regarding the sample size calculation are provided in [13]. For comparison, we also present the sample size required to obtain 90% marginal power for each outcome. For all calculations, we have used the Bonferroni method to account for multiple comparisons. We provide the sample sizes required to analyse two, three and four outcomes in Tables 3, 4 and 5, respectively. In Table 5, the top line provides an example sample size calculation for four outcomes where there is a small standardised effect size

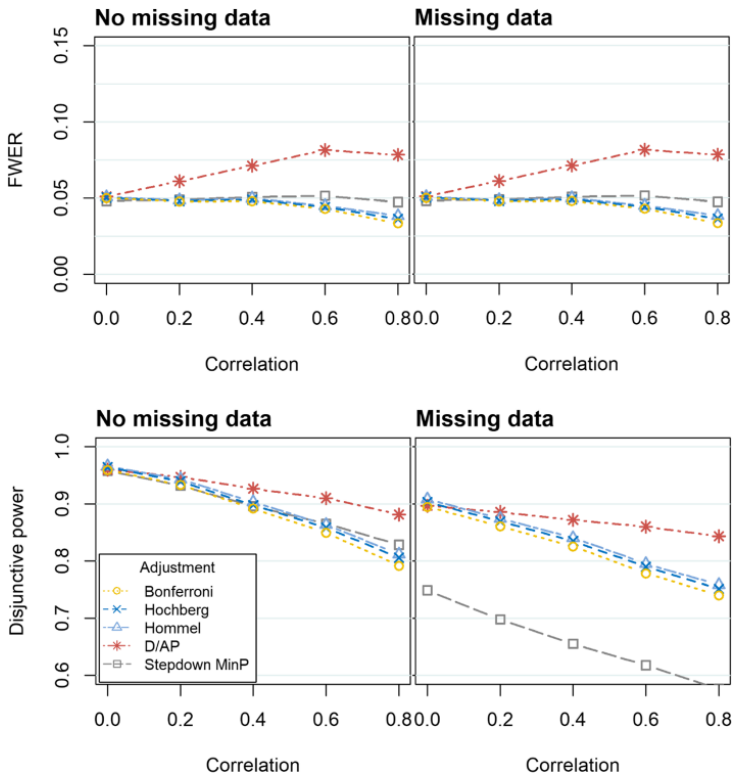


Fig. 2 FWER (top) and disjunctive power (bottom) obtained when evaluating three continuous outcomes using a variety of methods to control the FWER. In the left hand graphs, there are no missing data. In the right hand graphs, the missing data are missing completely at random, with 15% missing in one outcome and 25% missing in the other two outcomes ('Missing data') The graphs display various degrees of correlation between the outcomes, ranging from $\rho = 0$ to $\rho = 0.8$. The Monte Carlo standard errors (MCSE) were similar across all methods. When there was no missing data, the MCSE was between 0.001–0.004 for the disjunctive power and 0.002–0.004 for the FWER. In the missing data scenario, the MCSE was between 0.001–0.004 for the disjunctive power and between 0.001–0.004 for the FWER

for all four outcomes ($\Delta = 0.2$). When there is weak pairwise correlation between all four outcomes ($\rho = 0.2$), 325 participants would be required into each arm to obtain 90% disjunctive power. As the pairwise correlation increases to $\rho = 0.8$ the required sample size increases to 529. The sample size required to obtain 90% marginal for each outcome in this scenario is 716 participants per trial arm. The number of participants required to obtain 90% marginal power is greater than the number of participants required to obtain 90% disjunctive power. Thus the required sample size varies considerably depending on whether marginal or disjunctive power is used. The smallest of the sample sizes required to obtain the desired marginal power is the required sample size to achieve 90% disjunctive power if the outcomes are perfectly correlated ($\rho = 1$) [37].

Discussion

When using multiple primary outcomes in RCTs it is important to control the FWER for confirmatory phase III trials. One approach to do this is to adjust the *p*-values produced by each statistical test for each outcome. Additionally, some of the outcomes are likely to have missing values, consequently this needs to be considered when choosing an appropriate method to adjust the *p*-values.

Statistical analysis

We found that all methods investigated, except the D/AP, controlled the FWER. This agrees with the results previously reported in [19]. The Stepdown minP performed best in terms of FWER, but the R package used to implement the method uses listwise deletion

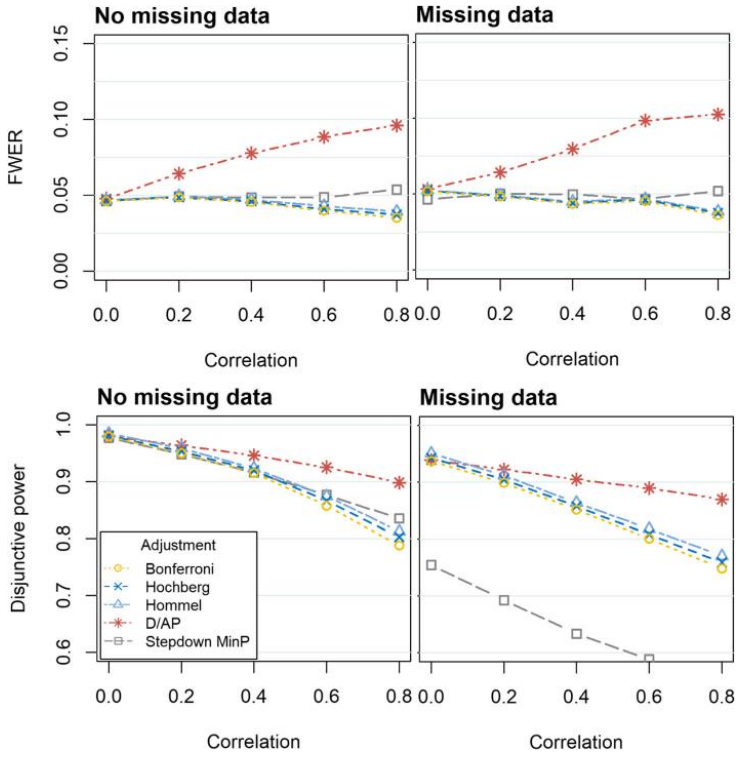


Fig. 3 FWER (top) and disjunctive power (bottom) obtained when evaluating four continuous outcomes using a variety of methods to control the FWER. In the left hand graphs, there are no missing data. In the right hand graphs, the missing data are missing completely at random, with 15% missing in two outcomes and 25% missing in the other two outcomes ('Missing data'). The graphs display various degrees of correlation between the outcomes, ranging from $\rho = 0$ to $\rho = 0.8$. The Monte Carlo standard errors (MCSE) were similar across all methods. When there was no missing data, the MCSE was between 0.001–0.004 for the disjunctive power and 0.002–0.004 for the FWER. In the missing data scenario, the MCSE was between 0.001–0.004 for the disjunctive power and between 0.001–0.004 for the FWER

removing participants with at least one missing value before the analysis resulting in a loss of power. The validity of this approach depends on how the method is implemented and the extent of the missing data.

We recommend that the Hommel method is used to control FWER when the distributional assumptions are met, as it provides slightly more disjunctive power than the Bonferroni and Holm methods. The distributional assumption associated with the Hommel method is not restrictive and is met in many multiplicity problems arising in clinical trials [22]. Even when the data followed a skewed distribution, the Hommel method performed well, showing it may be used to analyse a variety of outcomes, including those with a skewed distribution.

Given the availability of the software packages to implement the more powerful approaches, there is little reason to use the less powerful methods, such as Holm method. For example, the Hommel method can easily be implemented in R or SAS. Even though it is not currently available in Stata or SPSS, the p-values can be copied across and adjusted in R. However, if the assumptions cannot be met, the simpler Holm method could be used.

When the intervention effect size varied across the outcomes, we found that the differences in disjunctive power between the methods were less pronounced. It appeared that the outcome with the largest effect size 'dominated' the disjunctive power. When the sample size is based on the disjunctive power, the outcomes

Table 1 Marginal (individual) power obtained for each outcome, when analysing two (top), three (middle) or four (bottom) continuous outcomes using a variety of methods to control the FWER

Pairwise correlation between outcomes	None	Bonferroni	Holm	Hochberg	Hommel	Stepdown minP
Two outcomes						
0	80.9	72.4	78.5	79.2	79.2	78.2
0.2	80.6	71.8	77.8	78.6	78.6	77.7
0.4	80.0	71.3	76.6	77.7	77.7	76.7
0.6	80.0	71.0	76.0	77.4	77.4	76.7
0.8	80.3	71.3	75.6	77.4	77.4	77.2
Three outcomes						
0	80.2	65.9	75.2	76.7	76.8	75.5
0.2	80.5	66.4	75.0	76.6	76.7	75.3
0.4	80.2	65.7	73.8	75.4	75.6	73.2
0.6	80.0	65.7	73.3	75.0	75.2	73.8
0.8	80.0	65.9	72.2	74.6	74.8	76.1
Four outcomes						
0	80.5	62.3	73.2	75.0	75.2	72.7
0.2	80.4	62.3	72.6	74.4	74.8	72.2
0.4	80.6	62.4	72.1	74.1	74.4	72.2
0.6	80.3	62.0	70.7	73.1	73.5	72.3
0.8	80.3	61.9	69.7	73.2	73.6	73.5

D/AP method was not examined due to the poor performance observed when exploring FWER
There was no missing data in any of the outcomes. The tables display various degrees of correlation between the outcomes, ranging from no correlation ($\rho = 0.0$) to strong correlation ($\rho = 0.8$)

with the largest effect size would have high marginal power, whereas the outcome with the smallest effect size would have low marginal power – much below the overall desired level of power. It follows that when investigators are looking for an intervention effect for at least one outcome, it is unlikely that they will see an intervention effect on the outcomes with the smaller effect sizes without seeing an intervention effect on the outcomes with the largest effect size. Consequently, in this scenario, it may be advisable to pick the outcome(s) with the largest effect size as the primary outcome(s) and treat the other outcomes as secondary outcomes, however, this decision will need to account for the relative clinical importance of the outcomes. Alternatively, when the intervention effect size varies across the outcomes, investigators may wish to consider 'alpha spending' in which the total alpha (usually 0.05) is distributed or 'spent' across the M analyses.

We appreciate that in practice the choice of the adjustment method may also depend on other factors, such as the availability of simultaneous confidence intervals and unbiased estimates. It is standard practice to report the 95% confidence intervals alongside point estimates and *p*-values. When using multiple primary outcomes, it may be necessary to adjust the confidence interval so that it corresponds to the *p*-values adjusted for multiplicity. The confidence interval may be easily adjusted when using Bonferroni or Holm adjustments, using the R function "AdjustCIs" in the package "Mediana" [38]. However, it is not straightforward to adjust the confidence interval when using the Hochberg and Hommel. Consequently, the confidence intervals reported may not align with the *p*-values when these adjustments are used. As stated in the European Medical Agency (EMA) guidelines, in this instance, the conclusions should be based on the p-values and not the confidence intervals [3]. If confidence intervals that correspond to the chosen multiplicity adjustment are not available or are difficult to derive, then the EMA guidelines advise that simple but conservative confidence intervals are used, such as those based on Bonferroni correction [3].

The statistical analysis plan of a trial should clearly describe how the outcomes will be tested including which adjustment method, if any, will be used [39].

Our review of trials with multiple outcomes showed that majority of the trials analysed the outcomes separately without any adjustments for multiple comparisons [4]. Where adjustment methods were used, only the

Table 2 The percentage of simulations in which an intervention effect was observed for neither outcome, one outcome or both outcomes when analysing two outcomes, using a variety of methods to control the FWER

Method	Pairwise correlation between outcomes	Number of outcomes an intervention effect was observed on		
		0	1	2
Bonferroni	0	16.1	48.4	35.5
	0.2	18.6	43.2	38.2
	0.4	20.6	37.7	41.7
	0.6	23.4	32.7	43.9
	0.8	26.3	26.3	47.5
Holm	0	16.1	35.6	48.3
	0.2	18.6	31.0	50.4
	0.4	20.6	26.4	53.0
	0.6	23.4	22.0	54.6
	0.8	26.3	16.0	57.7
Hochberg	0	15.1	35.6	49.4
	0.2	17.6	31.0	51.5
	0.4	19.3	26.4	54.3
	0.6	22.0	22.0	56.0
	0.8	24.8	16.1	59.1
Hommel	0	15.1	35.6	49.4
	0.2	17.6	31.0	51.5
	0.4	19.3	26.4	54.3
	0.6	22.0	22.0	56.0
	0.8	24.8	16.1	59.1
Stepdown minP	0.0	23.7	37.5	38.8
	0.2	25.6	33.6	40.8
	0.4	29.6	27.1	43.4
	0.6	32.2	20.2	47.6
	0.8	33.8	13.8	52.4

In these simulations there was missing data in the outcomes (15% in one outcome and 25% in the other outcome). The tables display various degrees of correlation between the outcomes, ranging from no correlation ($\rho = 0.0$) to strong correlation ($\rho = 0.8$)

most basic methods were used, possibly due to their ease of implementation. The Bonferroni method was the most commonly used method, although the Holm and Hochberg methods were also used. As a consequence, we focused on relatively simple techniques in this paper. However, more advanced approaches, such as graphical methods to control the FWER are available and described in Bretz et al. [40] and Bretz et al. [41] .

It is not necessary to control the FWER for all types of trial designs, for example, for trial designs with co-primary outcomes where all outcomes have to be declared statistically significant for the intervention to be deemed successful. The FDA guidelines state that in this scenario no adjustment needs to be made to control the FWER [39] and the 'conjunctive' power is used. We have not evaluated the conjunctive power as it is not relevant to the scenarios considered in this paper. The conjunctive power may be substantially reduced compared to the marginal power for each outcome [39] and is never larger than the marginal power [13]. The conjunctive power behaves in reverse to the disjunctive power in that as the correlation between the outcomes increases, the conjunctive power increases.

Additionally, multiplicity adjustments may not be necessary for early phase drug trials. However, it is generally accepted that adjustments to control the FWER are required in confirmatory studies, that is when the goal of the trial is the definitive proof of a predefined key hypothesis for the final decision making [42].

Sample size

When designing a clinical trial, it is important to calculate the sample size needed to detect a clinically important intervention effect. Usually the number of participants that can be recruited in a trial is restricted because of ethical, cost and time implications. The sample size

Table 3 Sample size required to obtain 90% disjunctive power and 90% marginal power when analysing two outcomes, after applying a Bonferroni correction

Standardised effect sizes for each of the 2 outcomes		Sample size required to obtain 90% DISJUNCTIVE power				Sample size required to obtain 90% MARGINAL power for each outcome	
		Correlation between outcomes					
Outcome 1	Outcome 2	0.2	0.4	0.6	0.8	Outcome 1	Outcome 2
0.2	0.2	402	436	475	522	622	622
0.2	0.3	237	251	264	274	622	278
0.2	0.4	145	150	154	156	622	157
0.2	0.5	96	98	99	100	622	101
0.3	0.3	179	194	211	232	278	278
0.3	0.4	126	135	144	152	278	157
0.3	0.5	89	93	97	99	278	101
0.4	0.4	101	109	119	131	157	157
0.4	0.5	78	84	90	96	157	101
0.5	0.5	65	70	76	84	101	101

Sample sizes provided are required per arm. A Bonferroni correction is applied for all calculations to account for the multiple comparisons

calculation for a trial is usually based on an appropriate statistical method which will be used for the primary analysis depending on the study design and objectives. The sample size can vary greatly depending on if the marginal power or overall disjunctive power is used highlighting the importance of calculating the sample size based on the trial objective. To account for multiplicity in the sample size calculation, we recommend that the Bonferroni adjustment is used. The Bonferroni adjustment can be applied easily within the sample size calculation using an analytical formula [39] and our simulation study showed that it maintains the FWER to an acceptable level for low to moderate correlation between the outcomes. Additionally, there is not much loss in power when using the Bonferroni adjustment, compared to the other methods, in the presence of missing data. In contrast, the other methods investigated in this paper are data driven and therefore it is not clear how these can be incorporated without prior data.

One approach that has previously been used to calculate the sample size for multiple primary outcomes, was to calculate the sample size based on the individual marginal powers for each outcome and to choose the maximum sample size for the trial [43]. This approach guarantees adequate marginal power for each individual test. However, this approach will overestimate the number of participants required if the investigators are interested in disjunctive power. Moreover, it may be problematic to achieve that sample size in trials where recruitment is a problem and may result in trials being closed down prematurely.

Table 4 Sample size per group, assuming three outcomes, 90% disjunctive power, after applying a Bonferroni correction

Standardised effect sizes for each of the 3 outcomes			Sample size required to obtain 90% DISJUNCTIVE power				Sample size required to obtain 90% MARGINAL power for each outcome		
			Correlation between outcomes						
Out.[a] 1	Out. 2	Out. 3	0.2	0.4	0.6	0.8	Out. 1	Out. 2	Out. 3
0.2	0.2	0.2	353	401	456	524	677	677	677
0.2	0.3	0.3	185	207	229	254	677	302	302
0.2	0.4	0.4	109	120	131	143	677	171	171
0.2	0.5	0.5	71	77	84	92	677	110	110
0.3	0.3	0.3	157	179	203	234	302	302	302
0.3	0.4	0.4	101	114	127	143	302	171	171
0.3	0.5	0.5	68	76	83	92	302	110	110
0.4	0.4	0.4	89	101	114	132	171	171	171
0.4	0.5	0.5	64	72	81	91	171	110	110
0.5	0.5	0.5	57	65	73	84	110	110	110

Sample sizes provided are required per arm. A Bonferroni correction is applied for all calculations to account for the multiple comparisons. Key: [a]'*Out*' Outcome

Table 5 Sample size per group, assuming four outcomes, 90% disjunctive power, after applying a Bonferroni correction

Standardised effect sizes for each of the 4 outcomes				Sample size required to obtain 90% DISJUNCTIVE power				Sample size required to obtain 90% MARGINAL power for each outcome			
				Correlation between outcomes							
Out.[a] 1	Out. 2	Out. 3	Out. 4	0.2	0.4	0.6	0.8	Out. 1	Out. 2	Out. 3	Out. 4
0.2	0.2	0.2	0.2	325	382	447	529	716	716	716	716
0.2	0.2	0.3	0.3	189	215	242	270	716	716	319	319
0.2	0.2	0.4	0.4	114	127	129	152	716	716	181	181
0.2	0.2	0.5	0.5	75	82	89	98	716	716	116	116
0.3	0.3	0.3	0.3	145	170	199	235	319	319	319	319
0.3	0.3	0.4	0.4	101	117	133	151	319	319	181	181
0.3	0.3	0.5	0.5	71	80	88	98	319	319	116	116
0.4	0.4	0.4	0.4	82	96	112	133	181	181	181	181
0.4	0.4	0.5	0.5	63	73	84	96	181	181	116	116
0.5	0.5	0.5	0.5	52	61	72	85	116	116	116	116

Sample sizes provided are required per arm. A Bonferroni correction is applied for all calculations to account for the multiple comparisons. Key: [a]'*Out*' Outcome

Finally, the sample size should be inflated to account for the expected amount of missing data.

Study extensions and limitations

In this paper, we only explored continuous outcomes. However, in RCTs binary outcomes or a combination of continuous and binary outcomes may be used. For two binary outcomes, the maximum possible pairwise correlation between the outcomes will be less than one in absolute magnitude [44] and therefore we would expect similar results but with less pronounced differences between methods for the strong correlations.

Additionally, we only explored global effects, that is either no interventions effect on any of the outcomes ($\beta_{1j} = 0$) or an intervention effect on all the outcomes ($\beta_{1j} \neq 0$). Global effects are most realistic when the strength of the correlation between the outcomes is moderate to strong. However, in practice a mixture of no effects and some intervention effects may be observed, especially when the strength of the correlation between the outcomes is weak.

Conclusions

To ensure that the FWER is controlled when analysing multiple primary outcomes in confirmatory randomised controlled trials, we recommend that the Hommel method is used in the analysis for optimal power, when the distributional assumptions are met. When designing the trial, the sample size should be calculated according to the trial objective. When specifying multiple primary outcomes, if considered appropriate, the disjunctive power could be used, which has smaller sample size requirements compared to that when using the individual marginal powers. The Bonferroni adjustment can be used in the sample size calculation to account for multiplicity.

Additional file

Additional file 1 Sample size calculation methodology. Varying the effect size across outcomes. Skewed data. (DOCX 1675 kb)

Abbreviations

CI: Confidence interval; D/AP: Dubey/Armitage-Parmar; FWER: Familywise error rate; MCAR: Missing completely at random; SE: Standard error

Acknowledgements

Not applicable.

Authors' contributions

VV, RO and GA conceived the concept of this study. VV carried out the simulations and drafted the manuscript. RO and GA critically reviewed and made substantial contributions to the manuscript. All authors approved the final manuscript.

Funding

Victoria Vickerstaff is supported by Marie Curie Core funding grant [MCCC-FCO-16-U], National Institute Health Research School of Primary Care Research Seedcorn funding grant and UCLH Biomedical Research Centre. Rumana Omar and Gareth Ambler's research work was undertaken at University College London Hospitals /University College London who received a proportion of funding from the United Kingdom Department of Health's National Institute for Health Research Biomedical Research Centres (NIHR BRC) funding scheme.

Availability of data and materials

The datasets analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Appendix 3

Received: 4 December 2018 Accepted: 21 May 2019
Published online: 21 June 2019

References

1. Teixeira-Pinto A, Siddique J, Gibbons R, Normand S-L. Statistical approaches to modeling multiple outcomes in psychiatric studies. Psychiatr Ann. 2009; 39(7):729.
2. De Los Reyes A, Kundey SMA, Wang M. The end of the primary outcome measure: a research agenda for constructing its replacement. Clin Psychol Rev. 2011;31(5):829–38.
3. European Medical Agency: Guideline on multiplicity issues in clinical trials.2017.
4. Vickerstaff V, Ambler G, King M, Nazareth I, Omar RZ. Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. Contemp Clin Trials. 2015;45:8–12.
5. Campbell AN, Nunes EV, Matthews AG, Stitzer M, Miele GM, Polsky D, Turrigiano E, Walters S, McClure EA, Kyle TL. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. Am J Psychiatr. 2014;171(6):683–90.
6. Middleton S, McElduff P, Ward J, Grimshaw JM, Dale S, D'Este C, Drury P, Griffiths R, Cheung NW, Quinn C. Implementation of evidence-based treatment protocols to manage fever, hyperglycaemia, and swallowing dysfunction in acute stroke (QASC): a cluster randomised controlled trial. Lancet. 2011;378(9804):1699–706.
7. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. J Res Educ Effectiveness. 2012;5(2):189–211.
8. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.
9. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–2.
10. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988;75(2):383–6.
11. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med. 1997; 16(22):2529–42.
12. Bretz F, Hothorn T, Westfall P. Multiple comparisons using R. Boca Raton: CRC Press; 2010.
13. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharm Stat. 2007;6(3):161–70.
14. Dmitrienko A, Tamhane AC, Bretz F. Multiple testing problems in pharmaceutical statistics. Boca Raton: CRC Press; 2009.
15. Chow S-C, Shao J, Wang H, Lokhnygina Y. Sample size calculations in clinical research. Boca Raton: Chapman and Hall/CRC; 2017.
16. Odekerken VJ, van Laar T, Staal MJ, Mosch A, Hoffmann CF, Nijssen PC, Beute GN, van Vugt JP, Lenders MW, Contarino MF. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. Lancet Neurol. 2012;12(1):37–44.
17. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol. 2014;14(1):118.
18. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402.
19. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF III, Butters MA. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology. 2009;23(2):255.
20. Lafaye de Micheaux P, Liquet B, Marque S, Riou J. Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. J Biopharm Stat. 2014;24(2):378–97.
21. Wright SP. Adjusted p-values for simultaneous inference. Biometrics. 1992; 48(4):1005–13.
22. Dmitrienko A, D'Agostino R. Traditional multiplicity adjustment methods in clinical trials. Stat Med. 2013;32(29):5172–218.
23. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment, vol. 279. New York: Wiley; 1993.
24. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test. 2003;12(1):1–77.
25. Reitmeir P, Wassmer G. Resampling-based methods for the analysis of multiple endpoints in clinical trials. Stat Med. 1999;18(24):3453–62.
26. Li D, Dye TD. Power and stability properties of resampling-based multiple testing procedures with applications to gene oncology studies. Comput Math Methods Med. 2013;2013:610297.
27. Capizzi T, Zhang J. Testing the hypothesis that matters for multiple primary endpoints. Drug Inf J. 1996;30(4):949–56.
28. Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the health technology assessment journal. Trials. 2018;19(1):544.
29. Thompson SG, Nixon RM. How sensitive are cost-effectiveness analyses to choice of parametric distributions? Med Decis Mak. 2005;25(4):416–23.
30. Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-Centre differences into cost-effectiveness evaluations. Health Econ. 2005;14(12):1217–29.
31. Beeken R, Leurent B, Vickerstaff V, Wilson R, Croker H, Morris S, Omar R, Nazareth I, Wardle J. A brief intervention for weight control based on habit-formation theory delivered through primary care: results from a randomised controlled trial. Int J Obes. 2017;41(2):246–54.
32. Osborn DP, Hardoon S, Omar RZ, Holt RI, King M, Larsen J, Marston L, Morris RW, Nazareth I, Walters K. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. JAMA Psychiatry. 2015;72(2):143–51.
33. Hassiotis A, Poppe M, Strydom A, Vickerstaff V, Hall IS, Crabtree J, Omar RZ, King M, Hunter R, Biswas A. Clinical outcomes of staff training in positive behaviour support to reduce challenging behaviour in adults with intellectual disability: cluster randomised controlled trial. Br J Psychiatry. 2018;212(3):161–8.
34. Killaspy H, Marston L, Green N, Harrison I, Lean M, Cook S, Mundy T, Craig T, Holloway F, Leavey G. Clinical effectiveness of a staff training intervention in mental health inpatient rehabilitation units designed to increase patients' engagement in activities (the rehabilitation effectiveness for activities for life [REAL] study): single-blind, cluster-randomised controlled trial. Lancet Psychiatry. 2015;2(1):38–48.
35. Kohl M, Kolampally S. mpe: multiple primary endpoints; 2017.
36. Scherer R. Samplesize: sample size calculation for various t-tests and Wilcoxon-Test; 2016.
37. Sozu T, Kanou T, Hamada C, Yoshimura I. Power and sample size calculations in clinical trials with multiple primary variables. Jpn J Biometrics. 2006;27(2):83–96.
38. Paux G, Dmitrienko A. Package 'Mediana': Clinical Trial Simulations. 1.0.7 ed; 2018.
39. Food, Administration D: Multiple endpoints in clinical trials guidance for industry. Food and Drug Administration Draft Guidance. Multiple endpoints in clincial trials guidnace for industry. Silver Springer. 2017.
40. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. Biom J. 2011;53(6):894–913.
41. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med. 2009;28(4):586–604.
42. Bender R, Lange S. Adjusting for multiple testing—when and how. J Clin Epidemiol. 2001;54(4):343–9.
43. Allen RP, Chen C, Garcia-Borreguero D, Polo O, DuBrava S, Miceli J, Knapp L, Winkelman JW. Comparison of pregabalin with pramipexole for restless legs syndrome. N Engl J Med. 2014;370(7):621–31.
44. Warner RM. Applied statistics: from bivariate through multivariate techniques: sage; 2008.

Publisher's Note

Appendix 4

Sample size calculation to obtain desired marginal and disjunctive power

This appendix contains additional background on the methodology for the sample size calculations described in the method section of the Chapter 4. The results described in this section are a concise summary of the relevant results that can be found in textbooks on sample size calculations (Machin et al., Sozu et al., 2015).

In all trials, the power requirements should match the clinical objective which should be pre-specified when designing the study and the sample size should be performed accordingly. In current practice sample size calculations often focus on the marginal power for each outcome. However, we may also be interested in the disjunctive power. In this appendix, we describe the sample size calculation required assuming that we are interested initially in maximising the marginal power and secondly the disjunctive power.

We assume that we have a two-arm trial in which there are M primary outcomes. We are interested in testing the null hypotheses H_j $(j = 1, \dots, M)$ that there is not an intervention effect on the corresponding outcomes. The test statistics z_j are used to test the null hypotheses H_j. Further suppose that there is an overall null hypothesis $H_0(M) = \bigcap_{j=1}^{M} H_j$. Under this overall hypothesis, the joint test statistic $(z_1, \dots, z_M)$ has an M-variate distribution.

A4.1 Sample size calculation for marginal power

We use the marginal power when we are interested in the power to detect an intervention effect on a nominated outcome. The desired marginal power may be specified for each outcome. In this case, we test the hypothesis null H_j that there is not an intervention effect on the corresponding outcome.

To estimate the sample size we used an unpaired Student's t-test and we assumed equal variances. Suppose we wish to detect a standardised effect size Δ_j , then for significance level α, and power $1 - \beta$, the number of subjects per group is given by:

$$n = 2\frac{2\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{\Delta_j^2}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the $\left(1-\frac{\alpha}{2}\right)$ and $(1-\beta)$ quantiles of the standard normal distribution respectively. Further details on the sample size calculation based on the marginal power can be found in the textbook 'Sample Size Tables for Clinical Studies' by Machin et al. (2018). In the article, we calculated the required sample size using the R package "samplesize".

A4.2 Sample size calculation for disjunctive power

We use the disjunctive power when we are interested in testing the overall null hypothesis ($H_0(M)$) that there is no difference between intervention groups for all M outcomes. The overall alternative hypothesis ($H_1(M)$) is that there is an intervention effect on at least one of the M outcomes. We assume that the variances are known. For the standardised effect size Δ_j, $j = 1, \ldots, M$, the overall disjunctive power is

$$\begin{aligned} 1-\beta &= P\left[\bigcup_{j=1}^{M}\left\{Z_j > z_{1-\frac{\alpha}{2}}\right\} \,\middle|\, H_1(M)\right] \\ &= 1 - P\left[\bigcap_{j=1}^{M}\left\{Z_j \le z_{1-\frac{\alpha}{2}}\right\} \,\middle|\, H_1(M)\right] \\ &= 1 - P\left[\bigcap_{j=1}^{M}\left\{Z_j^* \le c_j^*\right\} \,\middle|\, H_1(M)\right] \end{aligned}$$

where $Z_j^* = Z_j - \sqrt{jn}\,\Delta_j$ and $c_j^* = z_{1-\frac{\alpha}{2}} - \sqrt{jn}\,\Delta_j$ and n is the number of subjects per group. The vector of test statistics $(Z_1^*, \ldots, Z_j^*)$ is distributed as an m-variate normal distribution $N_M(\mathbf{0}, \boldsymbol{\rho_z})$ where the off diagonal element of $\boldsymbol{\rho_z}$ is given by $\rho^{jj'}$. The disjunctive power is calculated by using the cumulative distribution function of the m-variate normal distribution. The sample size is the smallest integer required to achieve the desired overall power of 1 − β at the significance level of α. Further details regarding this sample size calculation are provided in a textbook by Sozu et al. (2015).

In the body of the thesis, I calculated the sample size for a pre-specified disjunctive power using the R package “mpe”, in particular I used the command “atleast.one.endpoint”. The function can be used to computer the sample size for continuous multiple primary outcomes where a significant difference for at least one outcome is expected.

Appendix 5

Methods to adjust for multiple comparisons in the analysis of randomised controlled trials with multiple primary outcomes which have varying effect sizes or are skewed.

In this appendix, I examine methods to adjust for multiple comparisons in additional scenarios. I begin by varying the effect size across the outcomes. I then simulate data that has a skewed distribution.

A5.1 Varying the effect size across outcomes

The following results were obtained by assuming varying intervention effect sizes across continuous outcomes. When analysing two outcomes, I specified that the intervention effect sizes were 0.2 and 0.4 for the outcomes respectively. When analysing four outcomes, I specified that the intervention effect sizes were 0.1, 0.2, 0.3 and 0.4 for each of the outcomes respectively.

Figure A5.1 Disjunctive power obtained when evaluating two continuous outcomes using a variety of methods to control the FWER.

In the left-hand graph, there are no missing data. In the right-hand graph, the missing data are missing completely at random, with 15% missing in the first outcome and 25% missing in the second outcome ('Missing data MCAR'). The graphs display various degrees of correlation between the outcomes, range from $\boldsymbol{\rho = 0}$ to $\boldsymbol{\rho = 0.8}$.

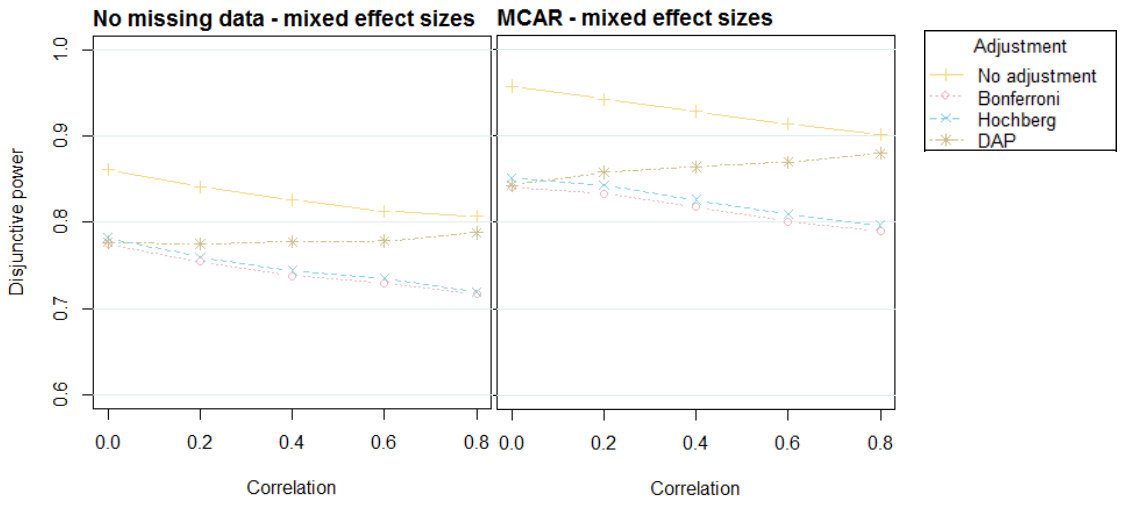
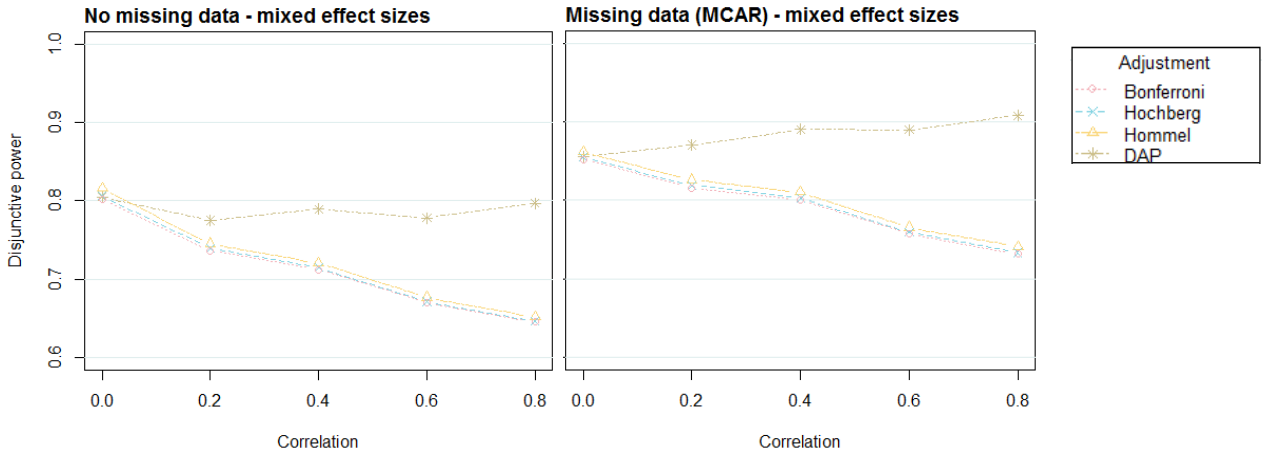


Figure A5.2 Disjunctive power obtained when evaluating four continuous outcomes using a variety of methods to control the FWER.

In the left-hand graph, there are no missing data. In the right-hand graph, the missing data are missing completely at random, with 15% missing in two outcomes and 25% missing in the other two outcomes ('Missing data MCAR'). The graphs display various degrees of correlation between the outcomes, range from $\boldsymbol{\rho = 0}$ to $\boldsymbol{\rho = 0.8}$.



*The Monte Carlo standard errors (MCSE) were similar for all methods. When there were no missing data, the MCSE was between 0.002-0.004 for the disjunctive power.

A5.2 Skewed data

I investigated the effect of skewed data by generating the outcome data (with equal intervention effect sizes) using a gamma distribution shape parameter = 2 and scale parameter =2. One iteration of the data are shown below to demonstrate the distribution of the data.

Figure A4.3 One iteration of the data drawn to investigate the effect of skewed data. Distribution of outcome 1 is shown on the left and the distribution of outcome 2 is shown on the right.

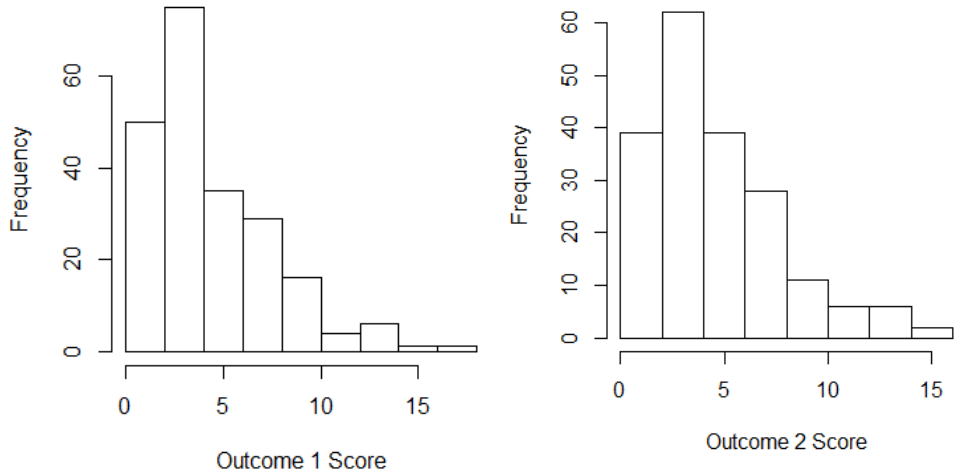
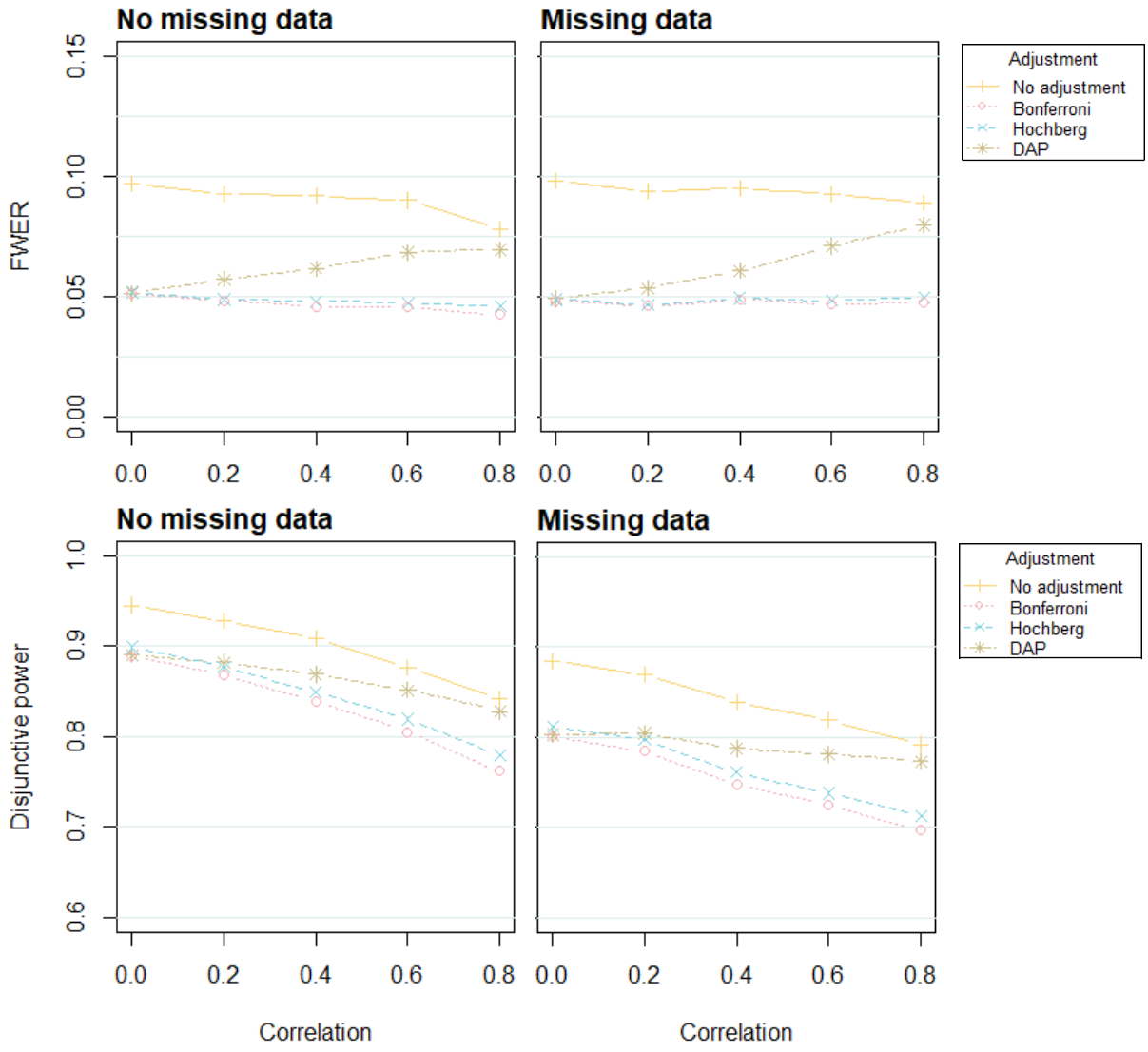


Figure A5.4 FWER (top) and disjunctive power (bottom) obtained when evaluating two continuous outcomes which have a skewed distribution, using a variety of methods to control the FWER.

In the left-hand graphs, there are no missing data. In the right-hand graphs, the missing data are missing completely at random, with 15% missing in the first outcome and 25% missing in the second outcome. The graphs display various degrees of correlation between the outcomes, range from $\boldsymbol{\rho = 0}$ to $\boldsymbol{\rho = 0.8}$.



Appendix 6

Bias and empirical standard errors obtained when using multivariate and univariate approaches to analyse outcomes and the data are MCAR and MAR

In this appendix, I provide additional simulation results to compare the multivariate multilevel (MM) model and the latent variable (LV) model in comparison to univariate models with and without multiple imputation (MI+UV, respectively). I investigate scenarios which vary in types of missingness, percentage of missing data and degree of correlation between outcomes. The results from this appendix are referred to in Chapter 5.

The figures in this section show the estimated intervention effects obtained when using the different methods to analyse two outcomes using various scenarios. The tables that follow show the empirical standard errors of the estimated intervention effects obtained in the different scenarios.

In this section when discussing low level of missing data, there is 15% and 25% missing outcome data for outcome 1 and outcome 2 respectively. For high level of missing data, there is 30% and 50% missing outcome data for outcome 1 and outcome 2 respectively.

Figure A6.1 Bias in estimating the intervention effects obtained when simulating two continuous outcomes. The blue dots represent the average of the estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 1. The red dots represent the average of estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 2. The five dots (of each colour) clustered together represents different correlation between the outcomes from 0 (left) to 0.8 (right) in increments of 0.2. Each graph displayed has different level/type of missing data as indicated. The true intervention effect is represented by the black horizontal line.

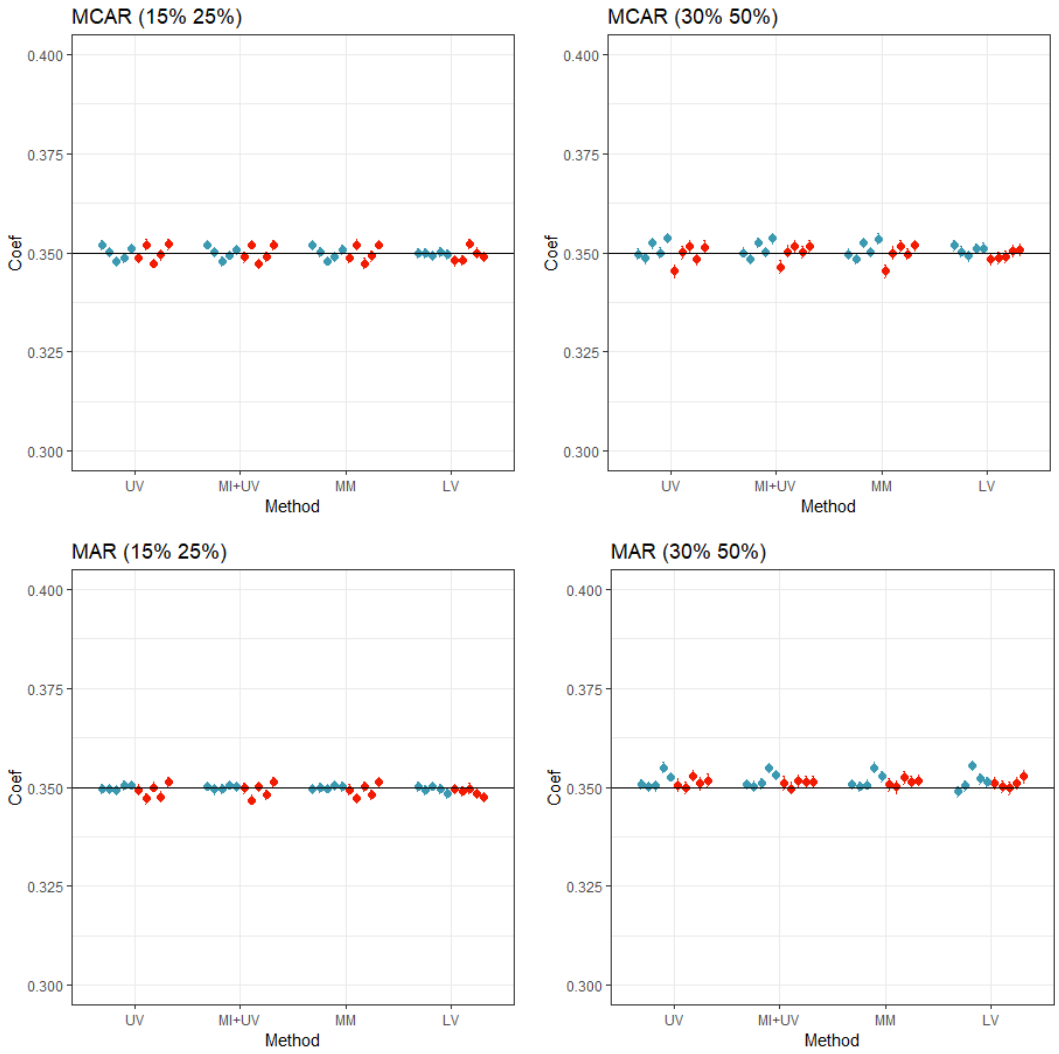


Figure A6.2 Bias in estimating the intervention effects obtained when simulating two binary outcomes. The blue dots represent the average of the estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 1. The red dots represent the average of estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 2. The five dots (of each colour) clustered together represents different correlation between the outcomes from 0 (left) to 0.8 (right) in increments of 0.2. Each graph displayed has different level/type of missing data as indicated. The true intervention effect is represented by the black horizontal line.

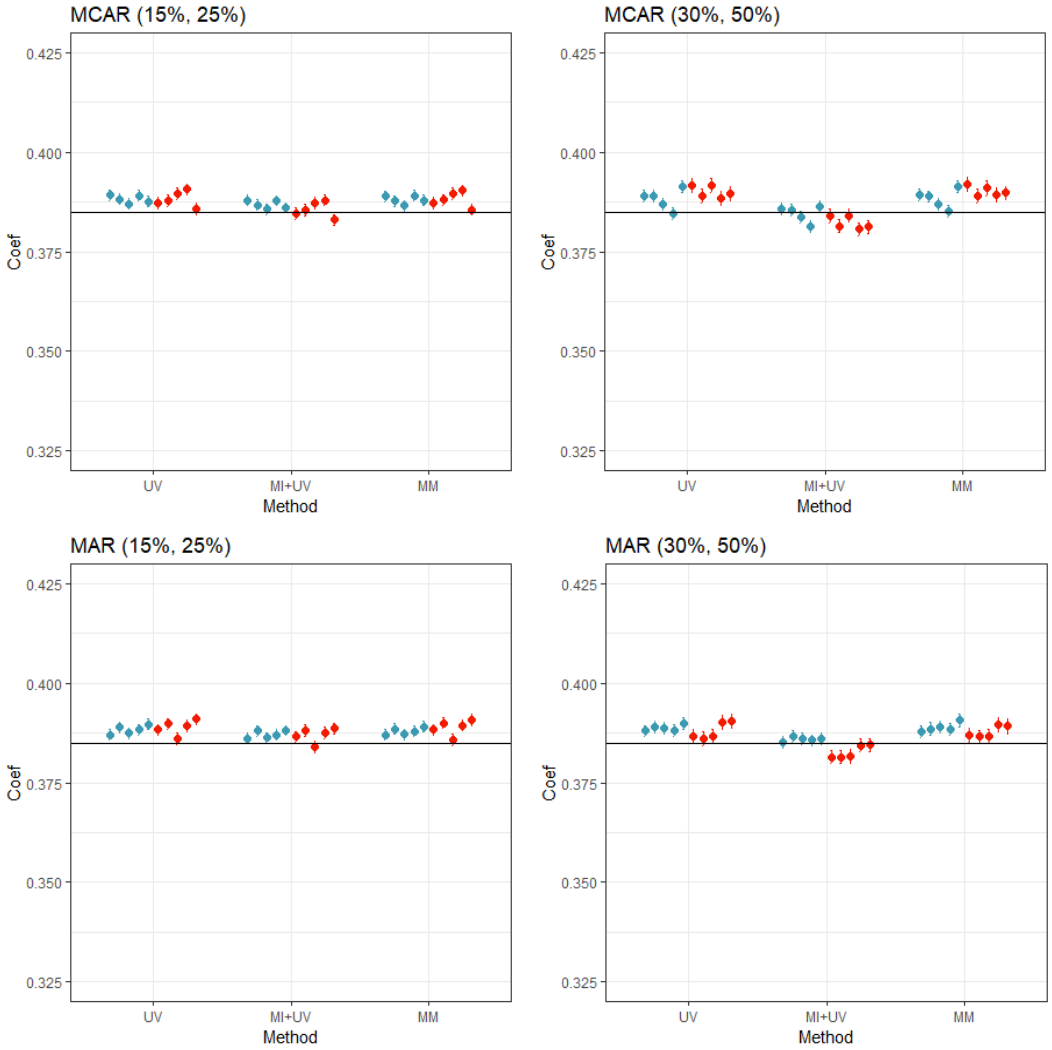


Figure A6.3 Bias in estimating the intervention effects obtained when simulating two mixed outcomes (outcome 1 is continuous and outcome 2 is binary). The blue dots represent the average of the estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 1. The red dots represent the average of estimated intervention effect ($\widehat{\boldsymbol{\beta}}$) for outcome 2. The five dots (of either colour) clustered together represents different correlation between the outcomes from 0 (left) to 0.8 (right) in increments of 0.2. Each graph displayed has different level/type of missing data, as indicatd. The true intervention effect is represented by the black horizontal line

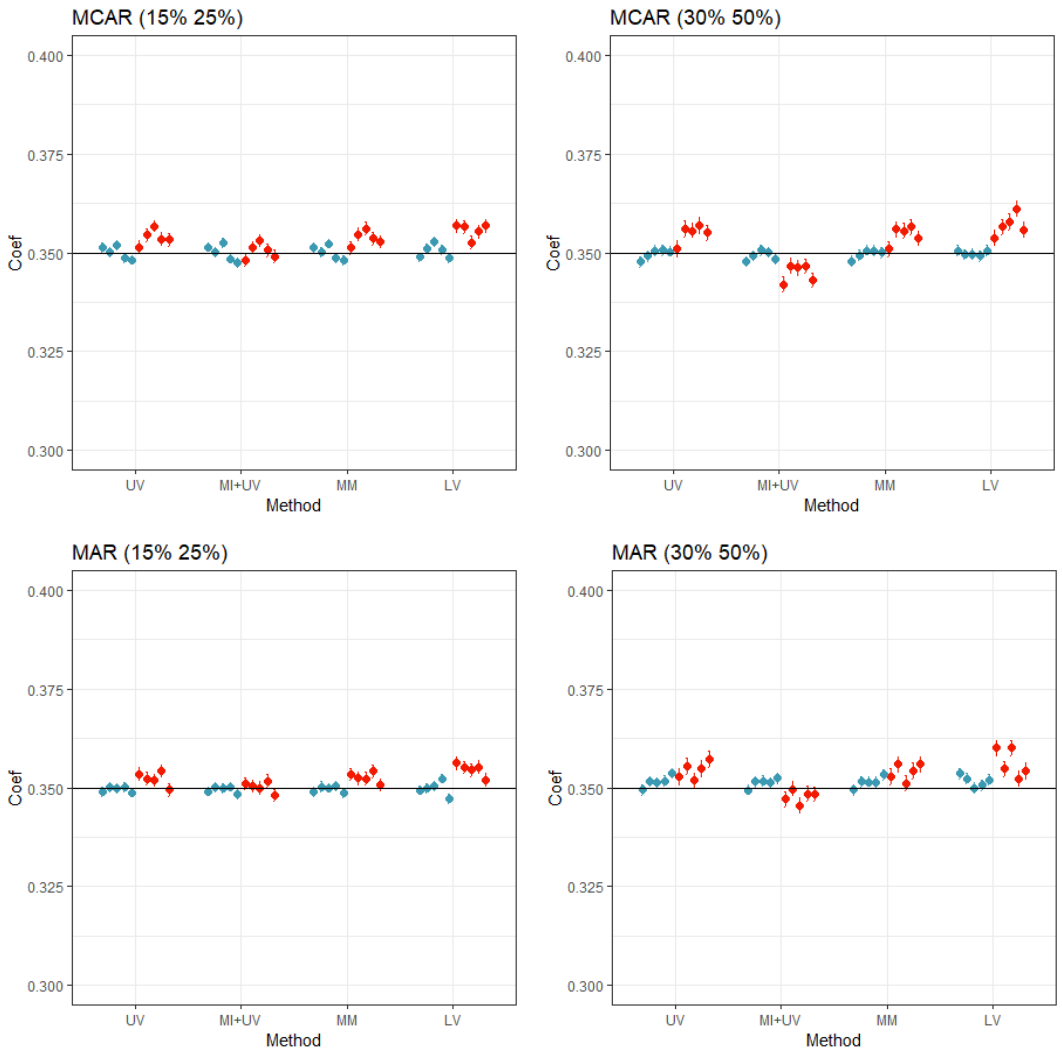


Table A6.1a Empirical standard error (EmpSE) of estimated intervention effect when evaluating two continuous outcomes

Type of Missingness ↓	% of missing values for each outcome ↓	ρ ↓	EmpSE of estimated intervention effect on outcome 1				EmpSE of estimated intervention effect on outcome 2			
		Method →	UV	MI + UV	MM	LV	UV	MI + UV	MM	LV
Complete	(0%, 0%)	0	0.123	-	0.123	0.123	0.124	-	0.124	0.124
		0.2	0.123	-	0.123	0.123	0.124	-	0.124	0.124
		0.4	0.125	-	0.125	0.125	0.124	-	0.124	0.124
		0.6	0.124	-	0.124	0.124	0.123	-	0.123	0.123
		0.8	0.125	-	0.125	0.128	0.124	-	0.124	0.128
MCAR	(15%, 25%)	0	0.134	0.137	0.134	0.134	0.145	0.153	0.145	0.145
		0.2	0.134	0.135	0.134	0.134	0.143	0.145	0.142	0.142
		0.4	0.135	0.134	0.134	0.134	0.143	0.142	0.141	0.141
		0.6	0.136	0.134	0.133	0.133	0.145	0.141	0.140	0.140
		0.8	0.133	0.130	0.129	0.131	0.143	0.135	0.134	0.137
MCAR	(30%, 50%)	0	0.148	0.156	0.148	0.148	0.177	0.197	0.177	0.177
		0.2	0.148	0.153	0.148	0.148	0.176	0.187	0.175	0.175
		0.4	0.148	0.148	0.147	0.147	0.177	0.175	0.171	0.171
		0.6	0.149	0.147	0.146	0.146	0.175	0.168	0.165	0.165
		0.8	0.147	0.143	0.141	0.142	0.175	0.157	0.154	0.155
MAR	(15%, 25%)	0	0.133	0.136	0.133	0.133	0.145	0.153	0.145	0.145
		0.2	0.134	0.135	0.134	0.134	0.143	0.146	0.143	0.143
		0.4	0.136	0.136	0.135	0.135	0.143	0.142	0.141	0.141
		0.6	0.134	0.132	0.132	0.132	0.143	0.139	0.138	0.138
		0.8	0.135	0.132	0.132	0.133	0.144	0.136	0.135	0.137
MAR	(30%, 50%)	0	0.147	0.155	0.147	0.147	0.178	0.199	0.178	0.178
		0.2	0.148	0.153	0.148	0.148	0.181	0.193	0.180	0.180
		0.4	0.148	0.149	0.146	0.146	0.180	0.179	0.175	0.175
		0.6	0.150	0.150	0.148	0.148	0.180	0.173	0.169	0.169
		0.8	0.148	0.143	0.142	0.143	0.180	0.160	0.157	0.158

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; $\rho *$ = correlation between outcomes.

Table A6.1b Empirical standard error (EmpSE) of estimated intervention effect when evaluating two binary outcomes

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	EmpSE of estimated intervention effect on outcome 1			EmpSE of estimated intervention effect on outcome 2		
		Method →	UV	MI + UV	MM	UV	MI + UV	MM
Complete	(0%, 0%)	0	0.138	-	0.141	0.137	-	0.137
		0.2	0.138	-	0.141	0.140	-	0.140
		0.4	0.137	-	0.140	0.139	-	0.139
		0.6	0.138	-	0.141	0.139	-	0.139
		0.8	0.139	-	0.142	0.138	-	0.139
MCAR	(15%, 25%)	0	0.150	0.150	0.151	0.162	0.161	0.162
		0.2	0.150	0.150	0.151	0.160	0.159	0.160
		0.4	0.150	0.149	0.151	0.159	0.158	0.158
		0.6	0.150	0.149	0.150	0.160	0.157	0.157
		0.8	0.149	0.147	0.148	0.161	0.153	0.155
MCAR	(30%, 50%)	0	0.166	0.165	0.167	0.198	0.196	0.199
		0.2	0.166	0.165	0.167	0.197	0.194	0.203
		0.4	0.166	0.164	0.166	0.195	0.190	0.194
		0.6	0.167	0.165	0.167	0.199	0.190	0.193
		0.8	0.165	0.161	0.163	0.199	0.184	0.188
MAR	(15%, 25%)	0	0.149	0.149	0.150	0.159	0.158	0.159
		0.2	0.151	0.151	0.152	0.159	0.159	0.160
		0.4	0.151	0.150	0.152	0.160	0.158	0.158
		0.6	0.149	0.148	0.149	0.162	0.159	0.160
		0.8	0.151	0.148	0.150	0.160	0.154	0.156
MAR	(30%, 50%)	0	0.165	0.164	0.166	0.200	0.197	0.201
		0.2	0.167	0.166	0.172	0.201	0.197	0.201
		0.4	0.167	0.166	0.167	0.202	0.196	0.199
		0.6	0.166	0.163	0.166	0.202	0.192	0.196
		0.8	0.168	0.163	0.170	0.204	0.186	0.191

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; ρ = correlation between outcomes.

Table 6.1c Empirical standard error (EmpSE) of estimated intervention effect when evaluating 'mixed' outcomes (one continuous and one binary)

Type of missingness ↓	% of missing values for each outcome ↓	ρ ↓	EmpSE of estimated intervention effect on outcome 1				EmpSE of estimated intervention effect on outcome 2			
		Method →	UV	MI + UV	MM	LV	UV	MI + UV	MM	LV
Complete	(0%, 0%)	0	0.124	-	0.160	0.160	0.124	-	0.160	0.161
		0.2	0.124	-	0.158	0.158	0.124	-	0.158	0.165
		0.4	0.125	-	0.158	0.158	0.125	-	0.158	0.186
		0.6	0.122	-	0.159	0.159	0.122	-	0.159	0.272
		0.8	0.124	-	0.160	0.160	0.124	-	0.160	0.504
MCAR	(15%,25%)	0	0.134	0.134	0.134	0.134	0.185	0.184	0.185	0.186
		0.2	0.134	0.135	0.134	0.134	0.182	0.181	0.182	0.190
		0.4	0.135	0.135	0.134	0.134	0.184	0.181	0.182	0.218
		0.6	0.134	0.133	0.133	0.133	0.185	0.179	0.180	0.325
		0.8	0.136	0.134	0.134	0.134	0.181	0.172	0.174	0.687
MCAR	(30%,50%)	0	0.150	0.150	0.150	0.150	0.224	0.221	0.225	0.228
		0.2	0.148	0.149	0.148	0.148	0.226	0.221	0.226	0.237
		0.4	0.148	0.148	0.147	0.147	0.226	0.218	0.222	0.275
		0.6	0.148	0.146	0.146	0.146	0.225	0.211	0.216	0.389
		0.8	0.149	0.145	0.145	0.146	0.228	0.203	0.210	0.553
MAR	(15%,25%)	0	0.133	0.133	0.133	0.133	0.184	0.183	0.184	0.185
		0.2	0.135	0.135	0.135	0.135	0.183	0.182	0.183	0.190
		0.4	0.135	0.135	0.134	0.134	0.183	0.181	0.182	0.219
		0.6	0.133	0.132	0.132	0.132	0.182	0.176	0.177	0.313
		0.8	0.134	0.131	0.131	0.132	0.183	0.174	0.176	0.515
MAR	(30%,50%)	0	0.149	0.150	0.150	0.150	0.229	0.225	0.230	0.233
		0.2	0.150	0.151	0.150	0.150	0.231	0.225	0.230	0.243
		0.4	0.149	0.149	0.148	0.148	0.230	0.220	0.226	0.278
		0.6	0.148	0.146	0.146	0.146	0.232	0.216	0.222	0.389
		0.8	0.149	0.145	0.145	0.146	0.229	0.205	0.212	0.566

Key: MM = multivariate multilevel model; UV = univariate model; MI + UV = multiple imputation followed by univariate model; LV = Latent variable model; ρ = correlation between outcomes.

Appendix 7

The implementation of the multivariate multilevel model using Stata, R and MlwiN

This appendix contains details on how to implement the multivariate multilevel (MM) model and latent variable model. I recommend that the software `MLwiN' is used to implement the MM model (Rabash et al., 2009). This software is freely available for academics through the Bristol University website. MLwiN can be used via R and Stata using the *R2MlwiN* (Zhang et al., 2016b) and *runMLwiN* (Leckie and Charlton, 2013) packages, respectively. Alternatively, when analysing multiple outcomes that are all the same type (say all continuous outcomes) the multivariate multilevel model can been implemented using a standard multilevel model. Multilevel models can be implemented in most standard packages (e.g. using *mixed* in Stata or *lmer* in R).

I used Stata to implement the latent variable model and therefore recommend the use of the mixed command or GLLAMM package to implement this model depending on the outcome type. The latent variable model could also be fitted using Proc NLMIXED in SAS.

In the sections below, I provide coding to implement the MM model using MLwiN via Stata and R and the latent variable model in Stata. For the following coding examples I assume that the data are in wide format, that is there is one line of data for each participant. If the data are required in a different format, this is described in more detail below with the code. In the examples I have a dataset called $dataSim$. This contains two outcomes (Y_1, Y_2); a variable to label if a participant received the intervention or not (arm); and, a participant identifier number (pid).

A7.1 Coding to implement MM model using MlwiN in Stata

In this section I explain how to analyse a dataset with two outcomes using the MM model in Stata. I provide coding to analyse two continuous outcomes, two binary outcomes and a mixture of outcome types (one continuous and one binary outcome). Prior to performing the analysis, MLwiN must be installed and a variable which is kept constant at 1 must also be created.

To begin, install the "runmlwin" package.

```
. ssc install runmlwin
```

Change the global MLwiN file path to the local file directory of MLwiN

```
. global MLwiN_path "PATH"      // where PATH is the local file directory to MLwiN
. gen cons = 1
```

To analyse two continuous outcomes, type

```
. runmlwin (Y1 arm cons, eq(1)) (Y2 arm cons , eq(2))   ///
        level1(pid: (cons, eq(1))(cons, eq(2))  nopause
```

To analyse two binary outcomes type

```
 . runmlwin (Y1 arm cons, eq(1)) (Y2 arm cons , eq(2)) ///
           level1(pid: (cons, eq(1))(cons, eq(2)))     ///
           discrete(distribution(binomial binomial) ///
           link(logit) denom(cons cons)) nopause
```

To analyse one continuous outcome and one binary outcome type

```
 . runmlwin (Y1 arm cons, eq(1)) (Y2 arm cons , eq(2)) ///
           level1(pid: (cons, eq(1))(cons, eq(2)))     ///
           discrete(distribution(normal binomial)      ///
           link(logit) denom(cons cons)) nopause
```

To extract the fixed effects and the corresponding variances type

```
 . matrix b_results = e(b)
 . matrix var_results = e(V)
```

A7.2 MM model using MLwiN in R

In this section I explain how to analyse a dataset with two outcomes using the MM model in R. Once again, MLwiN must be install before running this analysis. When using R, any missing data should be entered as “NA”.

To begin, install the “R2MLwin” package.

```
> install.packages("R2MLwiN")
```

To analyse continuous outcomes, type

```
> mv.model <- runMLwiN(c(Y1, Y2) ~ 1 + arm + (1| pid), D=Multivariate Normal", data=dataSim)
```

To analyse binary outcomes, type

```
> mv.model <- runMLwiN( c(probit(Y1, cons), probit(Y2, cons)) ~ 1 + arm , D=c("Mixed", "Binomial", "Binomial"), data = dataSim, estoptions = list(EstM = 1)
```

and to analyse one continuous and one binary outcome type

```
> mv.model <- runMLwiN( c(Y1, probit(Y2, cons)) ~ 1 + arm + (1[1] | pid ), D=c("Mixed", "Normal", "Binomial"), data = dataSim)
```

To extract the fixed effects and the corresponding standard errors type

```
> mv.model@FV

> mv.model@FV.cov
```

When analysing binary outcomes, a logit link function may also be used. The MM model runs quicker when using R compared to when using Stata.

A7.3 Latent variable model in Stata

In this section, I explain how to fit the latent model described by McCulloch (2008a) in Stata using the add in GLLAMM module. To implement this model, it is necessary to convert the data into a 'long' format. The following code can be used to analyse two continuous outcomes.

```
. findit gllamm
. eq het : cons_1 cons_2
. eq load: cons_1 cons_2
. gllamm outcome arm1 arm2 cons_1 cons_2,        ///
                s(het) i(pid) eq(load) nocons allc
```

Even though this coding can be used, when using the latent variable model to analyse two continuous outcomes in Stata, it is recommended to use `mixed` instead of the `gllamm`. Both methods require the same data preparation into the long format. The `mixed` command runs much quicker and more accurate results are obtained (Rabe-Hesketh and Skrondal, 2008). Under the assumption of normality for the random effects and of the outcomes given random effects, the likelihood has a simple closed form which the `mixed`

command utilises. On the other hand, gllamm uses numerical integration which is much slower. As all the features that are available in `gllamm` are also available with the `mixed` command the `GLLAMM` producers encourage the use of mixed instead (Rabe-Hesketh and Skrondal, 2008).

The following code can be used to analyse two mixed outcomes (1 continuous outcome and 1 binary outcome):

```
. eq het : cons_1
. eq load: cons_1 cons_2
. constraint define 2 [pid1_1]cons_1=0.8
. gllamm outcome arm1 arm2 cons_1 cons_2,         ///
      s(het) i(pid) eq(load) nocons allc constraint(2) ///
      family(gauss bin) link(id probit)           ///
      fv(response) lv(response)
```

Note that in the model with mixed outcome types it is necessary to impose a restriction on some of the variances. In the model, I have restricted that the factor variance to 0.8. For a discussion of alternative constraints see Skrondal and Rabe-Hesketh, 2004, pp. 107-108.

Further details for `gllamm` can be found on their webpage http://www.gllamm.org.

A7.3 Other R packages

An alternative multivariate approach could be implemented using the SabreR package in R. This package is designed to run "multivariate generalised linear mixed models". The outcomes can take the form of binary, ordinal, count and linear events and the different types can be combined. Currently (June 2019) the package cannot handle missing data. This is a downfall for the package and therefore I have not investigated it further as one main advantage of using multivariate analysis is the fact it can handle missing data. Details about the package can be found: http://sabre.lancs.ac.uk/model_intro.html.The team at Swansea University, lead by Prof. Damon Berridge are currently updating this software. They hope to provide a package which can handle multiple outcomes of mixed type and allowing for missing data. Consequently, this could be a viable option for multivariate analysis soon.

Appendix 8

Monte Carlo standard errors of the bias of the estimated intervention effects, empirical standard errors and coverage of the 95% confidence intervals when data are MNAR

Appendix 8

Table A8.1 Monte Carlo standard errors of the estimated bias when data are missing not at random using two continuous outcomes (left) and two binary outcomes (right)

% missing ↓	Correlation between outcomes ↓	Continuous 2 outcomes						Binary 2 outcomes					
	Method →	UV		MI + UV		MM		UV		MI+UV		MM	
	Outcome # →	1	2	1	2	1	2	1	2	1	2	1	2
Low 0%, 15%	0	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0007	0.0006	0.0007
	0.2	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0007	0.0006	0.0007
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0007	0.0006	0.0007
	0.6	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0007	0.0006	0.0008
	0.8	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0007	0.0006	0.0007
Medium 0% 30%	0	0.0006	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.2	0.0005	0.0006	0.0005	0.0007	0.0005	0.0006	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.6	0.0005	0.0006	0.0005	0.0006	0.0005	0.0006	0.0006	0.0008	0.0006	0.0007	0.0006	0.0008
	0.8	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0008	0.0006	0.0007	0.0006	0.0010
High 0% 50%	0	0.0006	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0018	0.0006	0.0012	0.0006	0.0015
	0.2	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0019	0.0006	0.0012	0.0006	0.0015
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0006	0.0017	0.0006	0.0012	0.0006	0.0013
	0.6	0.0005	0.0005	0.0005	0.0006	0.0005	0.0005	0.0006	0.0017	0.0006	0.0013	0.0017	0.0012
	0.8	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005	0.0006	0.0018	0.0006	0.0013	0.0006	0.0010
High overlapping 30% 50%	0	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0008	0.0017	0.0008	0.0012	0.0008	0.0014
	0.2	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0008	0.0018	0.0008	0.0012	0.0008	0.0014
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0008	0.0016	0.0008	0.0012	0.0008	0.0013
	0.6	0.0006	0.0005	0.0006	0.0005	0.0006	0.0005	0.0008	0.0017	0.0008	0.0012	0.0008	0.0012
	0.8	0.0006	0.0006	0.0006	0.0005	0.0006	0.0005	0.0008	0.0019	0.0007	0.0013	0.0009	0.0012

Table A8.2 Monte Carlo standard errors of the estimated bias when data are missing not at random using two 'mixed' outcomes.

% missing ↓	Correlation between outcomes ↓	Mixed 2 outcomes					
	Method →	UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2
Low 0%, 15%	0	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.2	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.4	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.6	0.0006	0.0008	0.0006	0.0008	0.0006	0.0008
	0.8	0.0005	0.0008	0.0005	0.0007	0.0005	0.0007
Medium 0% 30%	0	0.0006	0.0009	0.0006	0.0009	0.0006	0.0009
	0.2	0.0006	0.0009	0.0006	0.0009	0.0006	0.0009
	0.4	0.0006	0.0009	0.0006	0.0009	0.0006	0.0009
	0.6	0.0006	0.0009	0.0006	0.0008	0.0006	0.0008
	0.8	0.0006	0.0009	0.0006	0.0008	0.0006	0.0008
High 0% 50%	0	0.0006	0.0031	0.0006	0.0013	0.0006	0.0024
	0.2	0.0005	0.0031	0.0005	0.0013	0.0005	0.0020
	0.4	0.0006	0.0031	0.0006	0.0014	0.0006	0.0016
	0.6	0.0006	0.0031	0.0006	0.0014	0.0006	0.0013
	0.8	0.0006	0.0030	0.0006	0.0013	0.0006	0.0010
High overlapping 30% 50%	0	0.0006	0.0031	0.0006	0.0013	0.0006	0.0019
	0.2	0.0006	0.0031	0.0006	0.0013	0.0006	0.0019
	0.4	0.0006	0.0032	0.0006	0.0013	0.0006	0.0017
	0.6	0.0006	0.0032	0.0006	0.0013	0.0006	0.0014
	0.8	0.0006	0.0031	0.0006	0.0013	0.0006	0.0012

Table A8.3 Monte Carlo standard errors of the estimated bias when data are missing not at random using four continuous outcomes.

% missing ↓	Correlation between outcomes ↓	Continuous 4 outcomes											
	Method →	UV				MI + UV				MM			
	Outcome # →	1	2	3	4	1	2	3	4	1	2	3	4
Low 0%, 0% 15% 15%	0	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.2	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.6	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.8	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
Medium 0% 0% 30% 30%	0	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006
	0.2	0.0005	0.0006	0.0007	0.0006	0.0005	0.0006	0.0007	0.0006	0.0005	0.0006	0.0006	0.0006
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.6	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006
	0.8	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
High 0% 0% 50% 50%	0	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.2	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.4	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
	0.6	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0006	0.0006	0.0005	0.0005
	0.8	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0006	0.0006	0.0005	0.0005
High overlapping 20% 30% 40% 50%	0	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	0.2	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006
	0.4	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0005	0.0006	0.0006	0.0006	0.0005
	0.6	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005
	0.8	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005

Table A8.4 Monte Carlo standard errors of the estimated bias when data are missing not at random using four mixed outcomes.

% missing ↓	Correlation between outcomes ↓	Mixed 4 outcomes											
	Method →	UV				MI + UV				MM			
	Outcome # →	1	2	3	4	1	2	3	4	1	2	3	4
Low 0%, 0% 15% 15%	0	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008
	0.2	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008
	0.4	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008
	0.6	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0007	0.0006	0.0006	0.0007	0.0008
	0.8	0.0005	0.0006	0.0007	0.0008	0.0005	0.0006	0.0007	0.0007	0.0005	0.0006	0.0007	0.0008
Medium 0% 0% 30% 30%	0	0.0006	0.0006	0.0007	0.0009	0.0006	0.0007	0.0007	0.0009	0.0006	0.0006	0.0007	0.0009
	0.2	0.0006	0.0006	0.0007	0.0009	0.0006	0.0007	0.0007	0.0009	0.0006	0.0006	0.0007	0.0009
	0.4	0.0006	0.0006	0.0007	0.0009	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0009
	0.6	0.0006	0.0006	0.0007	0.0009	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008
	0.8	0.0006	0.0006	0.0007	0.0009	0.0006	0.0006	0.0007	0.0008	0.0006	0.0006	0.0007	0.0008
High 0% 0% 50% 50%	0	0.0006	0.0005	0.0007	0.0032	0.0006	0.0006	0.0007	0.0012	0.0006	0.0006	0.0007	0.0024
	0.2	0.0006	0.0006	0.0007	0.0033	0.0006	0.0006	0.0007	0.0012	0.0006	0.0006	0.0007	0.0026
	0.4	0.0005	0.0006	0.0007	0.0033	0.0005	0.0006	0.0007	0.0012	0.0005	0.0005	0.0007	0.0015
	0.6	0.0006	0.0006	0.0007	0.0030	0.0005	0.0005	0.0007	0.0013	0.0006	0.0005	0.0007	0.0012
	0.8	0.0005	0.0005	0.0007	0.0031	0.0005	0.0005	0.0007	0.0012	0.0005	0.0005	0.0007	0.0010
High overlapping 20% 30% 40% 50%	0	0.0006	0.0006	0.0010	0.0032	0.0006	0.0006	0.0010	0.0012	0.0006	0.0006	0.0010	0.0020
	0.2	0.0006	0.0006	0.0010	0.0032	0.0006	0.0006	0.0010	0.0012	0.0006	0.0006	0.0010	0.0018
	0.4	0.0006	0.0006	0.0010	0.0032	0.0006	0.0006	0.0009	0.0012	0.0006	0.0006	0.0010	0.0015
	0.6	0.0006	0.0006	0.0010	0.0030	0.0006	0.0006	0.0009	0.0012	0.0006	0.0006	0.0009	0.0012
	0.8	0.0006	0.0006	0.0010	0.0031	0.0006	0.0006	0.0009	0.0011	0.0006	0.0006	0.0009	0.0010

Table A8.5 Empirical standard errors of the estimated intervention effect for two continuous (left) and two binary (right) outcomes.

% missing ↓	Correlation between outcomes ↓	Continuous 2 outcomes						Binary 2 outcomes					
	Method →	UV		MI + UV		MM		UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2	1	2	1	2	1	2
Low 0% 15%	0	0.124	0.134	0.124	0.138	0.124	0.134	0.137	0.151	0.137	0.150	0.137	0.151
	0.2	0.124	0.134	0.124	0.135	0.124	0.134	0.140	0.151	0.140	0.151	0.140	0.152
	0.4	0.125	0.134	0.125	0.134	0.125	0.133	0.139	0.150	0.139	0.149	0.140	0.150
	0.6	0.124	0.134	0.124	0.131	0.124	0.130	0.138	0.150	0.138	0.148	0.138	0.177
	0.8	0.124	0.134	0.124	0.129	0.124	0.128	0.139	0.151	0.139	0.146	0.139	0.149
Medium 0% 30%	0	0.124	0.143	0.124	0.157	0.124	0.143	0.138	0.171	0.138	0.170	0.138	0.171
	0.2	0.123	0.143	0.123	0.147	0.123	0.142	0.139	0.170	0.139	0.169	0.140	0.172
	0.4	0.125	0.142	0.125	0.142	0.125	0.139	0.139	0.171	0.139	0.169	0.139	0.169
	0.6	0.122	0.142	0.122	0.141	0.122	0.135	0.139	0.171	0.139	0.167	0.139	0.189
	0.8	0.124	0.141	0.124	0.137	0.124	0.130	0.138	0.170	0.138	0.162	0.139	0.214
High 0% 50%	0	0.125	0.123	0.125	0.146	0.125	0.123	0.139	0.397	0.139	0.261	0.140	0.337
	0.2	0.125	0.124	0.125	0.138	0.125	0.123	0.138	0.415	0.137	0.264	0.138	0.325
	0.4	0.124	0.123	0.124	0.127	0.124	0.122	0.138	0.380	0.138	0.275	0.139	0.296
	0.6	0.122	0.122	0.122	0.124	0.122	0.119	0.139	0.372	0.139	0.284	0.383	0.267
	0.8	0.124	0.122	0.124	0.123	0.124	0.117	0.138	0.395	0.138	0.301	0.139	0.235
High overlapping 30% 50%	0	0.140	0.124	0.147	0.136	0.140	0.125	0.172	0.387	0.171	0.263	0.174	0.318
	0.2	0.143	0.124	0.147	0.133	0.142	0.124	0.172	0.394	0.171	0.261	0.180	0.313
	0.4	0.141	0.124	0.143	0.128	0.140	0.123	0.169	0.365	0.168	0.267	0.180	0.297
	0.6	0.141	0.123	0.140	0.122	0.139	0.120	0.171	0.384	0.169	0.278	0.183	0.278
	0.8	0.141	0.124	0.138	0.119	0.136	0.120	0.170	0.417	0.168	0.292	0.191	0.263

Table A8.6 Empirical standard errors of the estimated intervention effect for two 'mixed' outcomes.

% missing ↓	Correlation between outcomes ↓	Mixed 2 outcomes					
	Method →	UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2
Low 0% 15%	0	0.126	0.172	0.126	0.171	0.126	0.172
	0.2	0.124	0.172	0.124	0.171	0.124	0.171
	0.4	0.124	0.171	0.124	0.169	0.124	0.170
	0.6	0.123	0.172	0.123	0.168	0.123	0.169
	0.8	0.122	0.172	0.122	0.165	0.122	0.166
Medium 0% 30%	0	0.124	0.196	0.124	0.194	0.124	0.195
	0.2	0.125	0.194	0.125	0.192	0.125	0.194
	0.4	0.124	0.196	0.124	0.192	0.124	0.191
	0.6	0.123	0.197	0.123	0.189	0.124	0.188
	0.8	0.124	0.198	0.124	0.180	0.124	0.180
High 0% 50%	0	0.123	0.697	0.123	0.288	0.123	0.534
	0.2	0.123	0.682	0.123	0.299	0.123	0.441
	0.4	0.124	0.699	0.123	0.305	0.124	0.353
	0.6	0.124	0.686	0.124	0.313	0.124	0.286
	0.8	0.125	0.676	0.124	0.293	0.125	0.233
High overlapping 30% 50%	0	0.142	0.697	0.143	0.285	0.143	0.415
	0.2	0.141	0.698	0.142	0.286	0.141	0.421
	0.4	0.141	0.711	0.141	0.289	0.141	0.379
	0.6	0.141	0.719	0.141	0.297	0.141	0.317
	0.8	0.143	0.692	0.143	0.287	0.143	0.269

Table A8.7 Mean square error of the estimated bias, when exploring two continuous (left) and two binary (right) outcomes.

% missing ↓	Correlation between outcomes ↓	Continuous 2 outcomes						Binary 2 outcomes					
	Method →	UV		MI + UV		MM		UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2	1	2	1	2	1	2
Low 0% 15%	0	0.015	0.018	0.015	0.019	0.015	0.018	0.019	0.024	0.019	0.024	0.019	0.024
	0.2	0.015	0.018	0.015	0.018	0.015	0.018	0.020	0.024	0.020	0.024	0.020	0.024
	0.4	0.016	0.018	0.016	0.018	0.016	0.018	0.019	0.024	0.019	0.023	0.020	0.024
	0.6	0.015	0.018	0.015	0.017	0.015	0.017	0.019	0.024	0.019	0.023	0.019	0.032
	0.8	0.015	0.018	0.015	0.017	0.015	0.016	0.019	0.024	0.019	0.022	0.019	0.023
Medium 0% 30%	0	0.015	0.021	0.015	0.026	0.015	0.021	0.019	0.050	0.019	0.050	0.019	0.051
	0.2	0.015	0.021	0.015	0.023	0.015	0.021	0.019	0.051	0.019	0.050	0.020	0.050
	0.4	0.016	0.021	0.016	0.021	0.016	0.020	0.019	0.050	0.019	0.046	0.019	0.046
	0.6	0.015	0.021	0.015	0.020	0.015	0.019	0.019	0.051	0.019	0.044	0.019	0.051
	0.8	0.015	0.021	0.015	0.019	0.015	0.017	0.019	0.050	0.019	0.036	0.019	0.055
High 0% 50%	0	0.016	0.048	0.016	0.054	0.016	0.048	0.019	1.202	0.019	0.927	0.019	1.141
	0.2	0.016	0.046	0.016	0.050	0.016	0.045	0.019	1.229	0.019	0.920	0.019	1.106
	0.4	0.015	0.047	0.015	0.047	0.015	0.042	0.019	1.193	0.019	0.892	0.019	0.975
	0.6	0.015	0.047	0.015	0.043	0.015	0.034	0.019	1.183	0.019	0.827	0.147	0.746
	0.8	0.015	0.046	0.015	0.037	0.015	0.023	0.019	1.213	0.019	0.703	0.019	0.405
High overlapping 30% 50%	0	0.021	0.047	0.023	0.050	0.021	0.047	0.050	1.202	0.050	0.932	0.051	1.134
	0.2	0.021	0.047	0.023	0.049	0.021	0.046	0.052	1.211	0.053	0.918	0.056	1.113
	0.4	0.021	0.047	0.022	0.048	0.021	0.044	0.051	1.182	0.054	0.898	0.058	1.036
	0.6	0.021	0.047	0.022	0.044	0.021	0.038	0.050	1.195	0.056	0.848	0.062	0.880
	0.8	0.021	0.047	0.021	0.039	0.020	0.029	0.051	1.245	0.063	0.774	0.077	0.648

Table A8.8 Mean square error of the estimated bias, when exploring two mixed outcomes.

% missing ↓	Correlation between outcomes ↓	Mixed 2 outcomes					
	Method →	UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2
Low 0% 15%	0	0.016	0.030	0.016	0.030	0.016	0.030
	0.2	0.015	0.031	0.015	0.030	0.015	0.030
	0.4	0.015	0.030	0.015	0.030	0.015	0.030
	0.6	0.015	0.031	0.015	0.029	0.015	0.029
	0.8	0.015	0.031	0.015	0.028	0.015	0.028
Medium 0% 30%	0	0.015	0.058	0.015	0.058	0.015	0.058
	0.2	0.016	0.058	0.016	0.056	0.016	0.057
	0.4	0.015	0.059	0.015	0.053	0.015	0.053
	0.6	0.015	0.060	0.015	0.048	0.015	0.048
	0.8	0.015	0.058	0.015	0.039	0.015	0.039
High 0% 50%	0	0.015	1.725	0.015	0.920	0.015	1.439
	0.2	0.015	1.686	0.015	0.894	0.015	1.258
	0.4	0.015	1.718	0.015	0.806	0.015	0.975
	0.6	0.015	1.692	0.015	0.655	0.015	0.630
	0.8	0.016	1.677	0.015	0.420	0.016	0.317
High overlapping 30% 50%	0	0.021	1.711	0.021	0.881	0.021	1.259
	0.2	0.021	1.725	0.021	0.864	0.021	1.249
	0.4	0.021	1.742	0.022	0.799	0.022	1.081
	0.6	0.021	1.776	0.023	0.696	0.024	0.808
	0.8	0.022	1.721	0.026	0.514	0.029	0.511

Table A8.9 Coverage of 95% confidence intervals for the estimated intervention effects, when exploring two continuous (left) and two binary (right) outcomes.

% missing ↓	Correlation between outcomes ↓	Continuous 2 outcomes						Binary 2 outcomes					
	Method →	UV		MI + UV		MM		UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2	1	2	1	2	1	2
Low 0% 15%	0	95.0	94.8	95.0	94.6	94.9	94.7	95.2	93.9	95.2	93.9	95.3	93.8
	0.2	94.9	94.6	94.9	94.5	94.8	94.6	94.7	94.1	94.7	94.1	94.8	93.9
	0.4	94.6	94.7	94.6	94.5	94.5	94.5	95.0	94.2	95.0	94.2	95.0	93.9
	0.6	94.9	94.9	94.9	94.6	94.8	94.8	95.3	94.2	95.3	94.3	95.5	94.1
	0.8	94.8	95.1	94.8	94.5	94.7	94.7	94.8	94.2	94.8	94.5	94.8	94.0
Medium 0% 30%	0	94.9	94.2	94.9	93.2	94.8	94.0	95.0	85.7	95.0	85.5	95.1	85.3
	0.2	95.2	94.3	95.2	94.2	95.1	94.1	94.8	84.9	94.8	85.3	94.9	84.9
	0.4	94.6	94.1	94.6	93.9	94.5	94.0	95.1	85.6	95.1	86.8	95.2	86.6
	0.6	94.8	94.3	94.8	93.9	94.7	94.3	95.0	85.0	95.0	87.7	95.1	87.5
	0.8	95.1	94.5	95.1	93.7	95.0	94.8	95.1	85.2	95.1	90.4	95.1	90.3
High 0% 50%	0	94.7	69.5	94.7	70.3	94.6	68.8	95.2	2.5	95.2	3.6	95.3	2.4
	0.2	94.8	70.5	94.8	71.7	94.7	70.8	95.1	2.5	95.2	3.8	95.2	2.6
	0.4	94.9	69.8	94.9	69.6	94.8	71.9	95.5	2.5	95.4	6.0	95.5	3.3
	0.6	95.1	69.6	95.1	66.1	95.0	77.4	94.7	2.2	94.7	9.4	94.8	5.7
	0.8	95.0	70.2	95.0	63.6	94.9	86.2	95.0	2.6	95.1	18.9	95.3	18.7
High overlapping 30% 50%	0	94.3	70.3	96.8	79.2	94.2	69.5	85.3	2.5	85.4	3.5	87.7	2.3
	0.2	94.0	70.2	96.4	77.8	93.8	70.1	84.5	2.4	84.2	3.7	86.5	2.2
	0.4	94.3	69.5	95.7	74.0	93.9	70.7	84.6	2.1	83.1	4.6	85.4	2.4
	0.6	94.1	69.4	94.4	69.5	93.6	74.5	85.2	2.7	82.5	7.7	84.2	3.9
	0.8	94.4	69.9	93.9	66.9	93.5	81.9	84.9	2.6	79.3	12.7	79.4	6.8

Table A8.10 Coverage of the 95% confidence intervals, when exploring two mixed outcomes.

% missing ↓	Correlation between outcomes ↓	Mixed 2 outcomes					
	Method →	UV		MI + UV		MM	
	Outcome # →	1	2	1	2	1	2
Low 0% 15%	0	94.8	94.4	94.8	94.5	94.7	94.5
	0.2	94.8	94.4	94.8	94.5	94.7	94.5
	0.4	95.1	94.4	95.1	94.5	95.0	94.4
	0.6	95.2	94.4	95.2	94.7	95.1	94.7
	0.8	95.0	94.3	95.0	94.7	94.8	94.6
Medium 0% 30%	0	95.0	87.7	95.0	87.9	95.0	87.8
	0.2	94.7	87.7	94.7	88.3	94.5	88.1
	0.4	94.8	87.7	94.8	89.2	94.7	88.8
	0.6	95.1	87.3	95.1	89.8	94.9	89.6
	0.8	94.8	88.0	94.8	92.6	94.7	92.2
High 0% 50%	0	94.8	8.1	94.8	10.3	94.7	7.5
	0.2	95.1	8.8	95.1	12.7	95.0	8.4
	0.4	94.9	8.7	94.9	18.3	94.9	10.1
	0.6	94.6	8.6	94.7	30.5	94.8	17.4
	0.8	94.8	8.2	94.9	53.1	95.3	35.9
High overlapping 30% 50%	0	94.1	8.8	93.1	11.6	93.9	7.7
	0.2	93.9	8.7	92.7	12.7	93.7	8.1
	0.4	94.0	8.7	92.5	16.1	93.3	8.7
	0.6	94.2	8.4	91.7	23.7	92.4	12.5
	0.8	94.1	8.6	90.0	38.3	90.2	22.0