

Multiple Imputation for Handling Missing Outcome Data

Thomas Richard Sullivan

BMa&CompSc (Hons)

BSocSc

School of Public Health

Faculty of Health Sciences

The University of Adelaide

Australia

A thesis submitted in fulfilment of the requirements for the degree of Doctor of
Philosophy, October 2017

Contents

Abstract.....	vi
Declaration.....	viii
Manuscripts contributing to this thesis.....	ix
Presentations arising out of this thesis	x
Acknowledgements	xi
Abbreviations	xiii
1. Introduction.....	1
1.1. Multiple imputation	1
1.2. Multiple imputation and missing outcome data	2
1.3. Thesis aim	5
1.4. Thesis outline	6
2. Missing data and multiple imputation.....	7
2.1. Missing data	7
2.1.1. Missing data mechanisms.....	8
2.1.2. Missing data patterns.....	10
2.1.3. Statistical approaches to handling missing data	11
2.2. Multiple imputation	12
2.2.1. The imputation model	13
2.2.2. Multiple imputation inference and Rubin's rules	15
2.2.3. Methods for conducting multiple imputation	15
3. Multiple imputation of missing outcomes and thesis aims.....	20
3.1. Multiple imputation, then deletion	20
3.2. Multiple imputation for estimating the relative risk.....	23

3.3.	Multiple imputation in randomised trials.....	26
3.4.	Multiple imputation in extended follow-up studies.....	30
3.5.	Methods for addressing thesis aims.....	33
4.	Multiple imputation, then deletion.....	34
4.1.	Preface.....	34
4.2.	Statement of authorship	35
4.3.	Article	37
4.3.1.	Abstract.....	37
4.3.2.	Introduction.....	37
4.3.3.	Methods	41
4.3.4.	Results	44
4.3.5.	Discussion.....	49
4.3.6.	Web appendix	52
4.4.	Additional discussion.....	54
5.	Multiple imputation for estimating the relative risk	58
5.1.	Preface.....	58
5.2.	Statement of authorship	59
5.3.	Article	61
5.3.1.	Abstract.....	61
5.3.2.	Introduction.....	62
5.3.3.	Methods	64
5.3.4.	Results	69
5.3.5.	Discussion.....	77
5.3.6.	Web appendix	82

5.4.	Additional discussion.....	108
6.	Multiple imputation in randomised trials	110
6.1.	Preface.....	110
6.2.	Statement of authorship	111
6.3.	Article	113
6.3.1.	Abstract.....	113
6.3.2.	Introduction.....	113
6.3.3.	Intention to treat and missing data	116
6.3.4.	Methods.....	117
6.3.5.	Missing data in a univariate outcome.....	119
6.3.6.	Missing data in a multivariate outcome	126
6.3.7.	Missing data in a baseline covariate.....	130
6.3.8.	Case study.....	135
6.3.9.	Discussion.....	139
6.3.10.	Web appendix	142
7.	Multiple imputation in extended follow-up studies	148
7.1.	Preface.....	148
7.2.	Statement of authorship	149
7.3.	Article	151
7.3.1.	Abstract.....	151
7.3.2.	Introduction.....	152
7.3.3.	Methods.....	156
7.3.4.	Results	158
7.3.5.	Discussion.....	164

7.3.6.	Web appendix	168
7.4.	Guidance on the use of multiple imputation	178
7.4.1.	Multiple imputation and eligibility restrictions	179
7.4.2.	Multiple imputation and separate consent processes	183
7.4.3.	Multiple imputation and other sources of missing data	184
7.4.4.	Inverse probability weighting and multiple imputation	185
7.4.5.	Case study.....	187
7.4.6.	Conclusions.....	191
8.	Summary and conclusions	193
8.1.	Key findings and contributions	193
8.1.1.	Thesis aim 1	193
8.1.2.	Thesis aim 2	194
8.1.3.	Thesis aim 3	195
8.1.4.	Thesis aim 4	196
8.2.	Limitations and future directions.....	198
8.3.	Concluding remarks.....	200
9.	References.....	202

Abstract

Background: Multiple imputation is a widely used approach to handling missing data. Despite a growing evidence base for its use, implementation in practical settings remains challenging. This thesis considers knowledge gaps in the application of multiple imputation for handling missing outcome data.

Research has shown that deleting observations with multiply imputed outcomes before analysis can be beneficial when imputation and analysis models are the same. However, it is unclear how this approach performs with auxiliary variables, which are often available in practice. Another challenge arises when the outcome of interest is binary. The use of log binomial regression to produce relative risks is common, yet standard methods for imputing binary outcomes involve logistic regression or a multivariate normal assumption. It is uncertain whether inconsistencies between imputation and analysis models in this setting lead to biased or inefficient estimation. Questions also remain concerning the utility of multiple imputation in randomised trials. Unlike observational studies, the key exposure in randomised trials (randomised group) is always observed and independent of covariates for adjustment. If extended follow-up beyond completion of a randomised trial is planned, there may be more missing outcome data than in the original trial, and the use of eligibility restrictions and separate consent processes for participation in extended follow-up may complicate the use of multiple imputation. Unfortunately little is known about the extent of missing outcome data in this setting.

Aims: Specific aims are to:

1. Evaluate the effect of deleting imputed outcomes prior to analysis in the presence of auxiliary variables;
2. Investigate the performance of multiple imputation when estimating the relative risk;

3. Assess the utility of multiple imputation in randomised trials;
4. Summarise the extent of missing outcome data and provide guidance on the implementation of multiple imputation in extended follow-up studies.

Methods: The performance of multiple imputation was evaluated using data simulation and application to a real clinical trial. To summarise the extent of missing outcome data in extended follow-up studies, a systematic review of published follow-up studies was undertaken.

Results: Deleting imputed outcomes prior to analysis can lead to bias when the imputation model contains auxiliary variables associated with missingness in the outcome. For relative risk estimation, standard multiple imputation methods introduce bias and tend to produce confidence intervals that are too wide. Multiple imputation performs well in randomised trials, but simpler unbiased alternative methods for handling missing data are often slightly more efficient. Missing outcome data are a considerable threat to the validity of conclusions from extended follow-up studies. Eligibility restrictions and separate consent processes for participation are commonly employed in this setting, making the implementation of multiple imputation more challenging.

Conclusions: This thesis demonstrates the pitfalls of deleting imputed outcomes prior to analysis, the need for new methods of imputation when estimating the relative risk, and the limitations of multiple imputation for handling missing outcome data in randomised trials and extended follow-up studies. These findings will help to guide researchers on the appropriate use of multiple imputation for handling missing outcome data.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed:

Thomas Sullivan (PhD Candidate)

Date: 10/10/2017

Manuscripts contributing to this thesis

- Bias and precision of the “multiple imputation, then deletion” method for dealing with missing outcome data. *American Journal of Epidemiology*, 2015; 182(6): 528-34.
- Multiple imputation for handling missing outcome data when estimating the relative risk. *BMC Medical Research Methodology*, 2017; 17(1): 134.
- Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical Methods in Medical Research*, 2016; electronic publication available on journal website ahead of print, doi 10.1177/0962280216683570.
- Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. *Clinical Trials*, 2017; 14(4): 387-95.

Presentations arising out of this thesis

- Evaluation of multiple imputation approaches for handling missing outcome data. International Society for Clinical Biostatistics International Conference. Utrecht, Netherlands, August 2015.
- Multiple imputation for missing outcome data: to delete or not to delete? School of Public Health Seminar Series. Adelaide, Australia, June 2015.
- Bias and precision of multiple imputation then deletion in studies with missing outcome data. Young Statisticians National Conference. Adelaide, Australia, February 2015.
- Statistical approaches for handling missing data in randomised controlled trials. School of Public Health Seminar Series. Adelaide, Australia, August 2014.

Acknowledgements

I would like to acknowledge the following people for their support during the course of this research.

To my supervisors Amy Salter, Phil Ryan and Kate Lee, thank you for all your help throughout my candidature. To my principal supervisor Amy, thank you for always finding time for me, for all your encouragement, and for keeping me focused throughout the PhD. To Phil, thank you for everything you have done for me, not just during this PhD, but throughout my career as a statistician. It has been a privilege to work under you. To Kate, thank you for giving me the opportunity to work on some fantastic projects during this PhD, and for all your support and advice. I have learnt a great deal about multiple imputation under your guidance.

To members of the Melbourne multiple imputation group, thank you for all your help in developing the ideas for this thesis. It has been great to be part of such a supportive group throughout my PhD.

To Ian White, thank you for all your help and encouragement with the multiple imputation in randomised trials paper. It was truly a pleasure to work with you.

To Carmel Collins and Maria Makrides, thank you for granting me permission on behalf of the DINO steering committee to use the DINO trial as a case study for this thesis. Thank you also for all the support you've given me outside this PhD.

To my family and friends, thank you for all your support throughout this journey. Special thanks go to my mum, dad and sister Laura, for always being there for me and encouraging me to do my best, and to my friend Wayne, for making me laugh and keeping me relatively sane; who knows how many thousands of zombies, aliens and the like were slain for the sake of this PhD.

To my wife Lisa, thank you so much for your constant love and support; I am not sure I could have done this without you. You have been amazingly patient with me, talked through all the issues, and been there through all the highs and the lows. It has been a heck of a journey over the last three years, and I feel very fortunate to have shared it all with you.

Lastly to my now one-year-old daughter Adeline, thank you for being such an inspiration to me. I am very lucky to have you in my life.

Abbreviations

ANOVA	Analysis of variance
ATE	Average treatment effect
CCA	Complete case analysis
DHA	Docosahexaenoic acid
DINO	Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcome in Preterm Infants
DMI	Deletion, then multiple imputation
FCS	Fully conditional specification
FFM	Fat free mass
HC	Head circumference
ICH	International Conference on Harmonization
IPW	Inverse probability weighting
IQR	Interquartile range
ITT	Intention to treat
LMM	Linear mixed model
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MID	Multiple imputation, then deletion
MNAR	Missing not at random
MVNI	Multivariate normal imputation
NICE-SUGAR	Normoglycemia in Intensive Care Evaluation - Survival Using Glucose Algorithm Regulation
OR	Odds ratio
RCT	Randomised controlled trial
RR	Relative risk
SD	Standard deviation
SE	Standard error

1. Introduction

Missing data are a widespread problem in medical research. Defined as values that are not available but would have been meaningful for analysis had they been observed (1), missing data can result in biased and/or inefficient parameter estimates if inadequately handled in the statistical analysis. The validity of any statistical approach for handling missing data depends on the process that led to the data being missing, termed the "missing data mechanism". Rubin (2) introduced three classes of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are said to be MCAR if the probability of missing data is unrelated to observed or unobserved data. Data are MAR if the probability of missing data is unrelated to unobserved data, conditional on observed data. Lastly, MNAR occurs when the probability of missing data depends on unobserved data, even after taking observed data into account. Although MCAR can be ruled out from the observed data, it is not possible to distinguish between MAR and MNAR without knowing the missing values. Hence any analysis in the presence of missing data relies on untestable assumptions about the missing data mechanism.

1.1. Multiple imputation

Introduced by Rubin (3, 4), multiple imputation (MI) is a flexible and increasingly popular statistical approach for handling missing data. The method involves two distinct stages. In the first stage, each missing value is replaced by multiple draws ($m > 1$) from the posterior predictive distribution of the missing data conditional on the observed data, resulting in m complete datasets. In the second stage, the m complete datasets are analysed identically using standard complete-data techniques, with resulting estimates combined across datasets using rules that account for the uncertainty due to missing data. Standard implementations of MI assume that data are MAR, although the method can also be applied under an assumption that data are MNAR. Provided the assumption about the missing data

mechanism is satisfied and models used for imputation and analysis are correctly specified, MI produces consistent and asymptotically efficient parameter estimates (4).

A key task in applying MI is the specification of an appropriate method for generating the imputed datasets. To avoid bias, the model for imputing missing values should include all variables involved in the intended analysis in the appropriate functional form, accommodating non-linear and interaction terms as required (5). It can also be beneficial to include auxiliary variables, which are variables not involved in analysis models but added to the imputation model to improve estimation. Candidate auxiliary variables are correlates of analysis model variables that have missing data, correlates of missingness in those variables, or both (6). As well as decisions around the inclusion of variables in the imputation model, a general method for implementing MI must be chosen; a variety of methods are now available, each with their own strengths and weaknesses. In the case of missing outcome data, a choice must also be made about whether to retain or delete observations with imputed outcomes from imputed datasets. In the complex settings in which MI is typically applied, there is often no consensus in the literature to inform these decisions.

1.2. Multiple imputation and missing outcome data

Many of the challenges in implementing MI vary according to the nature of the missing data problem. The focus of this thesis is on the practical use of MI to handle missing data in outcome variables for analysis, which may or may not be accompanied by missing data in exposure variables. Missing outcome data are a common problem, particularly in randomised trials and observational studies involving longitudinal follow-up of participants. Indeed, in a recent systematic review on the use of MI in high impact medical journals, 72% of articles that stated which variables were included in the imputation model reported imputing missing outcome data (7).

In settings where both outcome and exposure variables are subject to missing data and interest concerns the estimation of regression coefficients from a generalised linear model, a popular alternative to the standard implementation of MI is the “multiple imputation, then deletion” (MID) method, proposed by von Hippel (8). MID entails imputing missing outcome and exposure values in the conventional manner, but then deleting observations with imputed outcomes prior to analysis. Provided imputation and analysis models are equivalent and correctly specified, MID can offer efficiency advantages over standard MI under a MAR assumption, particularly when the number of imputations is small (8). Another argument for MID is that it can help to minimise the bias introduced by a misspecified model for imputing missing outcomes, although this claim is yet to be supported by empirical evidence. A potential limitation of MID is its inability to incorporate information from auxiliary variables for the outcome. For a small number of imputations, von Hippel showed that the correlation between a single auxiliary variable and an incomplete outcome had to be fairly strong for standard MI to demonstrate efficiency advantages over MID (8). Unfortunately von Hippel only considered the efficiency of estimates in his evaluation of MID, ignoring the use of auxiliary variables for bias reduction. Importantly, it was unclear from this research whether MID could introduce bias when auxiliary variables associated with missing data in the outcome are included in the imputation model.

Another challenge arises when MI is applied in settings where the intended analysis has a functional form that is difficult to replicate in the imputation model. An important example of this problem is the use of MI for handling missing data in a binary outcome when the effect measure of interest is the relative risk. For missing data in both outcome and exposure variables, the two standard model-based methods of MI are fully conditional specification (FCS) (5, 9, 10), also known as chained equations or regression switching, and multivariate normal imputation (MVNI) (11). FCS involves specifying a series of univariate imputation models, one for each variable with missing data, with incomplete binary variables typically imputed using logistic regression. MVNI on the other hand assumes that all variables in the imputation model follow a multivariate

normal distribution. For analysis, the standard approach to estimating the relative risk is to fit a generalised linear model with a binomial error distribution and a log link, known as the log binomial model. This model assumes a different functional form for the relationship between the outcome and exposure variables than that involved in imputing outcomes using logistic regression in FCS, or under a multivariate normal assumption in MVNI. It is unknown whether inconsistencies between imputation and analysis models in this setting could lead to biased or inefficient estimation.

As well as challenges in implementing MI in general settings, this thesis considers the use of MI in randomised trials, where missing outcome data are often a major threat to the validity of group comparisons (12). Unlike observational studies, the key exposure in randomised trials (randomised group) is always observed and known to be independent of baseline covariates. In addition, missing data tend to be restricted to outcome variables in randomised trials, although baseline covariates may also be subject to missing data. Under these conditions, other methods for handling missing data may be preferable to MI. Should MI be adopted, an important consideration in handling missing outcome data is whether imputation should be carried out across all randomised participants or whether a separate but identical imputation model should be fitted to each randomised group. If subgroup analyses are of interest, interaction terms involving randomised group should be accounted for during the imputation process to avoid biasing interaction tests towards the null (5). In this case, performing imputation separately by randomised group can be appealing since it avoids the need to specify interaction terms in the imputation model (13-15). Often, though, subgroup analyses are not of interest, and it is unclear whether there is any merit in undertaking imputation separately by randomised group in this setting.

After the protocol defined completion of a randomised trial, investigators may choose to initiate an extended follow-up period to study longer-term impacts of the intervention (16). This type of study design is referred to as an “extended follow-up study” throughout this thesis. Missing outcome data can be a

considerable threat to the validity of group comparisons in this setting. As well as increased attrition over time, extended follow-up studies often involve additional eligibility restrictions and consent processes for inclusion in follow-up, which can further reduce participation rates. An important consideration in applying MI in extended follow-up studies is whether ineligible and non-consenting participants (where applicable) should be included in the imputation model. This decision could depend on the availability of auxiliary variables in the original trial to aid in the imputation of outcomes collected during extended follow-up, and the ability to satisfy an assumption about the missing data mechanism with the inclusion of ineligible and non-consenting participants in the analysis. The population for which the parameter of interest is defined (e.g. all randomised versus only those satisfying additional eligibility criteria) should also be taken into account when implementing MI in extended follow-up studies. Unfortunately discussion of these issues is lacking in the current literature and it is unclear how missing data are being handled in practice in this context.

1.3. Thesis aim

The overarching aim of this thesis is to address knowledge gaps in the practical application of MI for handling missing outcome data. Specific aims are to:

1. Contrast the performance of standard MI and MID when auxiliary variables associated with the incomplete outcome are included in the imputation model.
2. Evaluate the use of standard model-based methods of MI for handling missing outcome data when the analysis involves the estimation of relative risks.
3. Compare MI with alternative methods for handling missing data in randomised trials and explore the merits of imputing separately by randomised group in this context.

4. Summarise the extent and sources of missing outcome data in extended follow-up studies and provide guidance on the implementation of MI in this setting.

1.4. Thesis outline

The remainder of this thesis is structured as follows. Chapter 2 provides background material on missing data and MI in order to introduce key concepts and terminology that will be used throughout the thesis. In Chapter 3, literature on the practical use of MI for handling missing outcome data is reviewed to identify knowledge gaps and motivate the four specific aims of the thesis, as described above. A general description of the methods used to address the thesis aims is also provided in this chapter. The four thesis aims are then addressed in sequence through Chapters 4 to 7, with publications arising from the research included in each chapter. Lastly, a general discussion of results, limitations, suggestions for further research and concluding remarks are provided in Chapter 8.

2. Missing data and multiple imputation

In this chapter, introductory material on missing data and the MI method is presented in order to introduce concepts and terminology that will be used throughout the thesis. Section 2.1 provides a background to missing data, focusing on key concepts such as the missing data mechanism and the pattern of missing data. The MI procedure and its underlying assumptions are then described in Section 2.2.

2.1. Missing data

Missing data are defined as values that, for one reason or another, are not available, but would have been meaningful for analysis had they been observed (1). Despite the best efforts of researchers to collect complete data, missing data remain a common problem in medical research. In randomised trials, missing data can arise from participants withdrawing from the study, perhaps due to worsening of their disease, an adverse reaction to study procedures, or relocating to a new area. In longitudinal settings, participants may be lost to follow-up during the course of the study, preventing the collection of data at subsequent assessments. Individual measures could also be missing, possibly because measuring equipment was unavailable or not working correctly, a question was missed, or the participant skipped or refused a subtest. These are of course just a few of many reasons why missing data arise.

The major concern with missing data is the threat it poses to the validity of study findings. In most statistical packages, the default approach for handling missing data is to restrict the analysis to participants with complete data on all variables in the analysis model, which is known as a complete case analysis. There are two major statistical drawbacks with this approach. First, discarding information from partially observed cases can lead to a loss of precision in comparison to methods that can incorporate this information. Second, a complete case analysis can introduce bias if participants with complete data systematically differ from those

with incomplete data (14). Statistical drawbacks aside, discarding information from partially observed cases is wasteful of the resources devoted to collecting that information in the first instance (17, 18).

Given the problems associated with a complete case analysis, a broad range of statistical approaches have been developed to more adequately handle missing data. These approaches aim to provide valid inference in the presence of missing data, that is, to produce consistent estimates with standard errors and confidence intervals that appropriately account for sampling variability and uncertainty due to missing data (1). The validity of any statistical approach for handling missing data depends primarily on the process that led to the data being missing, referred to as the missing data mechanism, and the resulting pattern of missing data. These two important characteristics for describing missing data are introduced in the following sections.

2.1.1. Missing data mechanisms

Broadly, the missing data mechanism describes the process by which data become missing. Suppose in a study involving n participants that data are intended to be collected on p different variables, all of which will feature in the substantive analysis model. Let $Y = (Y_1, \dots, Y_p)$ be a matrix comprising of the complete data (i.e. what would be observed in the absence of missing data). Note that some of the p variables could be outcome variables and others exposure variables, although no distinction is made between variable types at this stage. If some observations are missing, Y can be partitioned into observed and missing components, denoted by Y_{obs} and Y_{miss} , respectively. Finally, let M represent a matrix of missing data indicators for Y , with $M_{ij} = 1$ if Y_{ij} is missing and 0 otherwise (for participant $i = 1$ to n , and variable $j = 1$ to p). The missing data mechanism is formally defined as the conditional probability distribution of the missing data indicators given the data that were intended to be collected, i.e. $P(M|Y)$. Alternatively, to emphasise that this distribution can depend on both

observed and missing values of Y , the missing data mechanism can also be expressed as $P(M|Y_{obs}, Y_{miss})$.

Following the framework introduced by Rubin (2), missing data mechanisms can be classified into three broad categories, as described below.

1. Missing completely at random (MCAR). The missing data mechanism is MCAR, or equivalently data are said to be MCAR, if the probability of missing data is unrelated to observed or unobserved data, i.e. $P(M|Y) = P(M)$. Under this mechanism, participants with complete data are representative of those with incomplete data, and so a complete case analysis will result in unbiased estimates. Assuming that data are MCAR is a strong assumption to make, however, and one that rarely holds in practice (1, 6, 19).
2. Missing at random (MAR). The missing data mechanism is MAR if the probability of missing data is unrelated to unobserved data, conditional on observed data, i.e. $P(M|Y) = P(M|Y_{obs})$. MAR is a considerably less restrictive and more realistic mechanism than MCAR. Should an analysis approach be valid under an assumption that data are MAR, it will also produce valid inference when data are MCAR.
3. Missing not at random (MNAR). The missing data mechanism is MNAR if the probability of missing data depends on unobserved data, even after taking observed data into account. Unlike MCAR and MAR, the missing data mechanism $P(M|Y)$ needs to be explicitly incorporated into the analysis to ensure valid inference when data are MNAR.

It is important to note that the missing data mechanism relates to both the data collected and the analysis undertaken. To illustrate, suppose that the probability of missing data in Y_1 is unrelated to unobserved data, conditional on observed data in Y_2 . Thus data in Y_1 would be considered to be MAR in an analysis incorporating

all observed data on Y_1 and Y_2 . However, should Y_2 be omitted from the analysis, data in Y_1 would instead be MNAR.

Provided missing data are unplanned rather than by design, as is assumed throughout this thesis, any analysis in the presence of missing data relies on untestable assumptions about the missing data mechanism. Although MCAR can be ruled out using observed data, for example by identifying predictors of missing data using logistic regression, it is not possible to distinguish between MAR and MNAR without knowing the values of the missing data. As a result, researchers are strongly encouraged to undertake sensitivity analyses to assess the robustness of findings to the assumption made about the missing data mechanism in the main analysis (1, 13, 17, 19-21).

2.1.2. Missing data patterns

The missing data pattern describes which values in the data are observed and which are missing, as defined by the matrix of missing data indicators M . Within the missing data literature, a distinction is often made between univariate, monotone and arbitrary patterns of missing data (14, 22). Data are said to be missing in a univariate pattern when missing data are confined to a single variable. A monotone pattern of missing data occurs when the p variables intended for collection can be ordered in such a way that, when Y_j is missing for a participant, then (Y_{j+1}, \dots, Y_p) are also missing. The monotone pattern tends to arise in longitudinal settings, where drop-out at a given time-point entails missing data on variables collected at subsequent assessments. Lastly, if data are missing in more than one variable, and these variables cannot be ordered to produce a monotone pattern, then they are said to be missing in an arbitrary pattern.

The missing data pattern is important to consider for two reasons. First, it determines which statistical approaches may be used, as some approaches for handling missing data are only applicable to certain missing data patterns. Second, the validity of some statistical approaches can depend both on the broad type of

missing data mechanism (i.e. MCAR, MAR, or MNAR) and the missing data pattern. For example, a complete case analysis can lead to bias and a loss of precision when data are MAR in an arbitrary pattern. However, when data are MAR in a univariate pattern in the outcome variable for analysis, a complete case analysis is both unbiased and fully efficient (18, 23, 24).

2.1.3. Statistical approaches to handling missing data

A range of alternative statistical approaches to complete case analysis have been developed to handle missing data, including single imputation methods, inverse probability weighting, likelihood-based methods, and MI. A very brief overview of these approaches is provided below (with the exception of MI, which is covered in Section 2.2).

Single imputation methods describe any procedure in which missing values are replaced with a single imputed value. Widely used methods include mean imputation, hot deck imputation, and the baseline or last observation carried forward for longitudinal data. Although single imputation methods are easy to understand and allow end users to proceed with the analysis as if all data were observed, their validity often depends on unrealistic assumptions about the missing data mechanism. For example, the last observation carried forward can introduce bias when outcome values change following the last observed measurement (13). Another concern with single imputation methods is that analyses are often conducted as if all data were observed (i.e. by employing variance estimators that are only appropriate for complete data), which can lead to overstated precision (13, 25).

Inverse probability weighting (IPW) is a modification of complete case analysis whereby complete cases are weighted in the analysis according to the inverse of the probability of being a complete case. Similar to the use of probability weights in the survey sample setting, the basic idea of IPW is to reweight complete observations so that they are representative of the entire sample. Provided the

model for assigning probability weights is correctly specified, IPW produces valid inference when data are MAR (26). However, in its standard implementation, IPW can be inefficient, as it discards information from partially observed cases. In addition, the method is difficult to apply to arbitrary missing data patterns (26).

Another option for valid inference under a MAR assumption is to use an estimation procedure based on the likelihood function of the observed data, for example maximum likelihood estimation or Bayesian posterior inference. Rubin (2) showed that the missing data mechanism $P(M|Y)$ drops out of the likelihood function, and hence can be ignored during estimation, provided that data are MAR (see (2) for technical details). Although this simplifies the estimation a great deal, the likelihood function may remain complex in the presence of incomplete observations (6), and so special computational techniques are often required (e.g. the expectation-maximisation algorithm for maximum likelihood estimation). As well as providing valid inference under a MAR assumption, likelihood-based methods are highly efficient (6). Despite their attractive statistical properties, likelihood-based methods can be difficult to implement in standard statistical software packages, particularly when incorporating information from auxiliary variables. Hence they are not as widely used as other approaches to handling missing data, most notably MI.

2.2. Multiple imputation

First introduced in the survey sample setting in 1978 (3), MI now has a very large bibliography in the medical research literature, including numerous review papers and texts (e.g. (4, 11, 27-29)). The popular approach involves two distinct stages. In the first stage, each missing value is replaced by $m > 1$ values drawn from an imputation model, a process which results in the generation of m complete datasets. The rationale for using $m > 1$ imputations is to propagate missing data uncertainty, a key shortcoming of single imputation methods (without appropriate variance correction). In the second stage of MI, the analysis of interest is conducted on each complete dataset, with results appropriately combined across

datasets to give a single MI estimate. Standard implementations of MI provide valid inference when data are MAR, although the approach can also be applied under MNAR mechanisms (4).

MI has many appealing features. Arguably the most important is the considerable flexibility of the method. As well as its ability to be validly implemented under both MAR and MNAR mechanisms for any pattern of missing data, MI enables end users to employ virtually any statistical technique appropriate for complete data, which makes it widely applicable. Another appealing feature of MI is its ability to incorporate information from auxiliary variables, which in the context of MI are defined as variables not involved in analysis models but added to the imputation model to improve estimation of the missing values. In practice, auxiliary variables can lead to noticeable gains in terms of bias reduction and increased efficiency (6). Finally, MI procedures are now widely available in most major statistical packages, including SAS, Stata and R.

2.2.1. The imputation model

The validity of MI depends primarily on how the imputed values are generated. Very broadly, an imputation method should, on average, provide reasonable predictions for the missing data and reflect all relevant sources of uncertainty. An imputation method that satisfies these conditions and leads to valid inference is said to be “proper” (see (4) for technical details). In practice, proper imputations tend to be created using Bayesian arguments. Under a MAR assumption, this entails drawing imputed values from the posterior predictive distribution of the missing data given the observed data. Let $P(Y|\theta)$ denote a parametric model for the complete data with population parameters θ , and $P(\theta)$ a prior distribution for θ (typically a non-informative prior is specified). Independently for $k = 1$ to m , Bayesian proper MI proceeds by first drawing $\theta^{(k)}$ from its posterior distribution $P(\theta|Y_{obs})$ (where $P(\theta|Y_{obs}) \propto P(\theta) \int P(Y_{obs}, Y_{miss}|\theta) dY_{miss}$, see (29)), then drawing imputed values for $Y_{miss}^{(k)}$ from its posterior predictive distribution $P(Y_{miss}|Y_{obs}, \theta^{(k)})$. As well as reflecting uncertainty in the imputed values due to

prediction error, this process also importantly acknowledges the uncertainty in the estimated model parameters.

To ensure valid inference, the method for generating imputed values should also preserve associations among variables that will be the subject of subsequent analyses. In particular, the imputation model should be “compatible” with the analysis model, where compatibility is defined in statistical terms as the existence of a joint model that contains both the imputation and analysis models as conditionals (30, 31). Effectively this means that the imputation model should include all variables in the intended analysis in the appropriate functional form, accommodating non-linear and interaction terms as required (5, 11, 30). To illustrate the problem of incompatibility, consider an analysis involving the simple linear regression of an incomplete exposure on a complete outcome. Failing to include the outcome in the imputation model would result in imputed values in the exposure bearing no relationship with the outcome, which in the subsequent analysis would lead to the regression coefficient being biased towards the null (excepting the case where there truly was no association between variables). Although compatibility is simple to achieve in some settings (as in the example above with the inclusion of the outcome in the imputation model), in others it can be quite complex, for example in analyses involving interaction or quadratic terms for incomplete exposures (5, 30, 32, 33), survival outcomes (30, 34), or fractional polynomials (35).

While the imputation model should include all variables in the intended analysis model, it should be noted that the converse is not required (36). As previously described, one of the appealing features of MI is the ability to include auxiliary variables in the imputation model to assist with the prediction of missing values. In this case incompatibility can be beneficial for estimation, both in terms of bias reduction and increased efficiency. Hence a general strategy when specifying an imputation model is that it should be at least as complex as the intended analysis model (6, 11).

2.2.2. Multiple imputation inference and Rubin's rules

Having multiply imputed the missing values from an appropriate imputation model, the analysis is then conducted on each of the m complete datasets. Let Q denote the population parameter of interest, \hat{Q}_k the estimate of Q from the k^{th} complete dataset and W_k the corresponding variance estimate for \hat{Q}_k . Note that parameter and variance estimates will differ across the m complete datasets due to differences in imputed values. Using Rubin's rules (4), the combined MI estimate \hat{Q}_{MI} is calculated as the mean of the m estimates, i.e. $\hat{Q}_{MI} = 1/m \sum_{k=1}^m \hat{Q}_k$. The estimated variance is given by $\text{var}(\hat{Q}_{MI}) = W + B(1 + 1/m)$, where $W = 1/m \sum_{k=1}^m W_k$ is the average within-imputation variance and $B = (m - 1)^{-1} \sum_{k=1}^m (\hat{Q}_k - \hat{Q}_{MI})^2$ the between imputation variance. Assuming Q is a scalar quantity, for example a regression coefficient, Wald-type significance tests and confidence intervals can be obtained using a t -distribution with $\nu = (m - 1)[1 + W/(1 + m^{-1})B]^2$ degrees of freedom. Wald-tests can also be extended to handle multivariate Q (11). Provided imputation and analysis models are correctly specified, estimates derived using Rubin's rules are both consistent and asymptotically efficient (4).

As outlined in White et al. (5), Rubin's rules can be used to combine any statistic that is an estimator of a population parameter, although in some cases a transformation may be required to ensure the statistic is approximately normally distributed (e.g. for an odds ratio or a standard deviation). In contrast, statistics that are not estimators of a population parameter, for example p-values, cannot be combined using Rubin's rules.

2.2.3. Methods for conducting multiple imputation

Following a decision to use MI, a method of imputation needs to be chosen. When data are missing in a single variable, a univariate imputation model can be applied with the model tailored to the variable being imputed, for example linear regression for a continuous variable or logistic regression for a binary variable. If

data are missing in a monotone pattern, imputations can be generated using a sequence of univariate imputation models, starting from the variable with the least missing data and proceeding through to the variable with the most missing data, conditioning at each stage on variables imputed earlier in the sequence (4). Again, the univariate imputation models can be tailored to the variables being imputed. When data are missing in an arbitrary pattern, as is typically the case in practice, variables need to be imputed simultaneously using iterative methods. The two main approaches for this are joint modelling and FCS, as detailed below.

Joint modelling

Joint modelling involves specifying a parametric joint model for Y . Available joint models include the multivariate normal model for continuous variables, the log-linear model for categorical variables and the general location model for a mixture of continuous and categorical variables. Due to the limited applicability of alternative joint models, the multivariate normal model is indisputably the most popular joint model in practice, with the MVNI procedure now available in most major statistical packages. First implemented by Schafer (11), MVNI uses a Markov chain Monte Carlo algorithm (known as data augmentation) for imputation. Initially, missing values are imputed based on assumed starting parameter values for the multivariate normal distribution. These are typically obtained from available data using the expectation-maximisation algorithm. Next, updated parameter values for the multivariate normal distribution are drawn from their posterior distribution based on the observed and imputed data. This iterative process of imputing missing values and drawing updated parameter values continues until these values converge to a stationary distribution (11, 36). Following these “burn-in” iterations, a set of imputed values is taken. In order to reduce dependence between imputations, additional iterations are performed before the next set of imputed values is obtained.

Due to the strong theoretical underpinnings of joint modelling and the ease of specifying imputation models, MVNI is an appealing method when multivariate

normality is reasonable. Clearly such an assumption is not always realistic, particularly when the imputation model contains skewed or binary variables. For skewed data, several authors have recommended transforming variables to better approximate normality prior to implementing MVNI (14, 17, 36). In contrast, others have found that transformations have little effect on estimation (37) or can even increase bias (38). Recent evidence suggests that the linearity of relationships between variables, rather than the skewness of marginal distributions, is the more important factor to consider before applying a transformation (39). In the case of binary variables, continuous imputed values obtained through MVNI often need to be classified into categories so that statistical methods appropriate for binary data can be applied (e.g. logistic regression for a binary outcome). Bernaards et al. (40) investigated several classification methods for binary variables and found that MVNI performed well in most settings, particularly when an adaptive rounding threshold¹ was used to classify imputed values. Several other authors have also reported good performance with MVNI for binary variables (11, 37, 41). Despite these and other promising findings, it remains difficult to make global statements about the robustness of MVNI to violations of multivariate normality, whether in the specific cases of skewed and binary variables or more generally.

Fully conditional specification

Rather than defining a full joint model for the data, FCS involves specifying a series of univariate models, one for each variable with missing data (5, 9, 10). The most appealing feature of FCS is the ability to tailor univariate models according to the distribution of the variable being imputed. For example, linear regression can be used to impute continuous variables, logistic regression to impute binary variables, and Poisson regression to impute count variables. Other appealing features of FCS include its ability to handle skip questions and, where appropriate,

¹ The adaptive rounding threshold is based on a normal approximation to the binomial distribution. Letting p_k denote the mean of a binary (0/1) variable in the k^{th} complete dataset and Φ^{-1} the quantile function of the normal distribution, the threshold is given by $p_k - \Phi^{-1}(p_k)\sqrt{p_k(1 - p_k)}$.

impose bounds on imputed values by drawing them from truncated predictive distributions (9). It is also easy to accommodate non-linear and interaction terms within the univariate imputation models.

For each variable with missing data, the FCS algorithm begins by replacing missing values with “place holder” values (42), often by way of mean imputation or simple random sampling from the observed data. The first variable with missing data, Y_1 say, is then regressed on other variables according to its specified univariate model, restricted to participants with observed values of Y_1 and using place holder values for other variables. Missing values in Y_1 are then replaced by simulated draws from their posterior predictive distribution (allowing for uncertainty in model parameters). The process is then repeated for the next variable with missing data, for example Y_2 , but this time incorporating imputed rather than place holder values for Y_1 into the estimation. This process continues until all incomplete variables have been imputed, which signals the completion of a “cycle”. Further cycles are then performed using the most recent imputed values in order to stabilise the distribution of parameters governing the imputations, after which a single imputed dataset is generated. Additional imputed datasets are obtained by independently repeating this process.

Despite being extremely flexible, FCS is not without limitations. One concern with the approach is the possibility of specifying univariate models where the conditional distributions implied do not correspond to a valid joint distribution. A potential consequence is that results could vary according to the ordering of regression models within the FCS procedure, which is clearly undesirable. Fortunately this issue seems to have little impact on results in practice (9, 10, 41, 43). Another drawback of FCS is the modelling effort required to generate imputed datasets. Since regression models need to be specified for each incomplete variable in the imputation model, FCS can become quite time consuming, particularly in datasets containing a large number of variables (41). Finally, like MVNI, FCS can produce biased results when assumptions of the

imputation model are incorrect, for example when skewed variables are imputed using linear regression models (41).

3. Multiple imputation of missing outcomes and thesis aims

Many of the practical challenges in applying MI vary according to the nature of the missing data problem. Having provided a general outline of the MI procedure in Chapter 2, the thesis now turns to specific challenges in applying MI when handling missing outcome data. In this chapter, literature on the practical use of MI for handling missing outcome data is reviewed to identify knowledge gaps and motivate the four specific aims of the thesis. A brief overview of the methods to be used to address the thesis aims is also provided.

3.1. Multiple imputation, then deletion

When missing data are evident in exposure as well as outcome variables, it is well known that the outcome should be included in the imputation model to avoid biasing associations towards the null (11, 14, 44). Whether imputed outcome values should be retained in subsequent analyses is less clear. In an influential article, von Hippel (8) proposed a modification to the standard implementation of MI that involved deleting imputed outcomes prior to analysis, an approach he termed “multiple imputation, then deletion” (MID). Suppose data are collected on an outcome variable Y and exposure variables $X = (X_1, \dots, X_p)$; note the change in notation from previous sections in order to now distinguish between outcomes and exposures. MID involves generating imputed values in the usual manner, that is, by including both Y and X in the imputation model, then discarding observations where Y has been imputed prior to analysis. The resulting modified datasets are then analysed as intended, with parameter estimates and standard errors combined across datasets using Rubin’s rules.

von Hippel advocated MID primarily on the grounds of efficiency. Provided that the imputation and analysis models are compatible and correctly specified, and assuming data are MAR, MID produces unbiased estimates of regression

coefficients with a greater precision than that of MI (8) (albeit efficiency gains tend to be minor unless the number of imputations is small). von Hippel also argued that MID can help to minimise the bias introduced by a misspecified model for imputing the missing outcomes, as problematic imputed values are removed from the analysis (8). Unfortunately this claimed advantage has not yet been supported by empirical evidence.

The rationale for MID is that, following imputation, observations with missing outcomes only add noise to the estimation procedure (8, 18). Although this assertion is valid when imputation and analysis models are compatible and correctly specified, it does not hold when the imputation model contains auxiliary variables for improving the prediction of missing outcome values. Importantly, while both MI and MID benefit equally from the inclusion of auxiliary variables for predicting missing values in X , only MI benefits from auxiliary variables for predicting missing values in Y (8). The additional information provided by auxiliary variables for Y may need to be fairly substantial, however, for MI to demonstrate efficiency advantages over MID. Using a simulation study, von Hippel found that MID was more efficient than MI, provided the correlation between a single completely observed auxiliary variable and an incomplete outcome did not exceed 0.7, 0.6, and 0.5 for 2, 5, and 10 imputations, respectively (8). Based on these results, and noting that auxiliary variables may be less useful in practice when they too are subject to missing data, von Hippel concluded that MID will typically be a superior strategy relative to MI.

There are two major limitations with von Hippel's investigation of MID and auxiliary variables. First, it is unclear whether MID would maintain similar efficiency advantages over MI with a larger number of imputations. Although early texts on MI suggest that 10 or fewer imputations are often adequate (11, 25, 36), more recent texts recommend performing many more (5, 45), and it is not uncommon for 50 or more imputations to be used in practice (7). Second, von Hippel only considered the use of auxiliary variables for efficiency gains, ignoring settings where they might be used instead for reducing bias. In developing high

quality imputations, numerous experts have recommended incorporating auxiliary variables that are associated with the incomplete variables to be imputed, the probability of missing data, or both (5, 6, 11, 46, 47). It is the inclusion of auxiliary variables related to the probability of missing data that is important for satisfying a MAR assumption and hence for minimising bias; such auxiliary variables were not considered in von Hippel's research. Consequently, it remains unclear whether MID could introduce bias not seen with a conventional MI approach when auxiliary variables associated with missing data in Y are included in the imputation model.

The extent to which auxiliary variables in MI can minimise bias has been studied extensively. In a landmark study, Collins et al. (6) demonstrated via simulation that failure to incorporate information from an auxiliary variable correlated with an incomplete variable and with missingness in that variable led to biased estimates of regression coefficients following MI. Conversely, adding several "junk" auxiliary variables to the imputation model (that were unrelated to the incomplete variable) did not adversely impact estimation. Based on these results, Collins et al. recommended researchers adopt inclusive strategies when selecting auxiliary variables for imputation models. Extending this work, Graham (48) observed that the magnitude of bias introduced by omitting an auxiliary variable for an incomplete variable depended on a number of factors: the proportion of missing data in the incomplete variable, the proportion of missing data in the auxiliary variable, the strength of the association between the auxiliary variable and the incomplete variable, and the strength of the association between the auxiliary variable and missingness in the incomplete variable. As a simple rule of thumb, Graham suggested that overlooking an auxiliary variable would lead to practically meaningful bias if its correlation with the incomplete variable and with missingness in the incomplete variable both exceeded 0.40. Similar results have been observed in other studies, with the effects of auxiliary variables ranging from little impact on inference (37, 49) through to noticeable reductions in bias and/or gains in efficiency (50-53). Of course, one should take care not to incorporate too

many auxiliary variables, as an overfit imputation model can result in unstable and biased estimates (50, 51).

Assuming a sensible imputation model, the literature indicates that auxiliary variables in MI at worst do little harm, and at best can be greatly beneficial for estimation. Consequently, a more thorough assessment of the relative merits of MID in settings where auxiliary variables for an incomplete outcome are available is of practical importance.

Thesis aim 1

The first aim of this thesis is to contrast the performance of MI and MID in settings where missing data are evident in both outcome and exposure variables, and where auxiliary variables associated with the outcome are included in the imputation model. Two types of auxiliary variables will be considered: those associated just with the outcome, and those associated with both the outcome and missingness in the outcome. The impact of using a large number of imputations on the comparison between MI and MID will also be explored. Thesis aim 1 is addressed in Chapter 4.

3.2. Multiple imputation for estimating the relative risk

As described previously, the imputation model should include all variables to be included in the intended analysis in the functional form required for analysis. Although considerable research has focused on the correct specification of imputation models when handling missing data restricted to exposure variables, for example in analyses involving interaction or quadratic terms for incomplete exposures (5, 30, 32, 33), less attention has been paid to challenges associated with imputing missing outcome data. A somewhat neglected problem is the use of MI for handling missing outcome data when the analysis involves a generalised linear model with a non-canonical link function. In this case, it may be difficult to replicate the functional form of the analysis model using standard model-based

methods of MI, particularly when exposure variables are also subject to missing data. An important example of this problem, and the focus of Chapter 5 of this thesis, is the use of MI for handling missing data in a binary outcome when estimating the relative risk.

The relative risk is a summary measure of effect for binary outcome data that is often of interest in medical research (54-57). Formally, the relative risk describes the probability (or risk) of experiencing an outcome of interest in one group relative to the probability in another. Letting p_1 and p_0 denote outcome probabilities in two groups for comparison, the relative risk is given by p_1/p_0 . Unlike the standard metric for binary outcome data, the odds ratio, defined as $[p_1/(1 - p_1)]/[p_0/(1 - p_0)]$, the relative risk is simple to interpret and has the attractive statistical property of being collapsible across covariate strata (58). Another appealing feature of the relative risk is that, for clustered and longitudinal data, marginal (population-averaged) and conditional (subject-specific) parameter values are identical (59).

The main drawback of the relative risk is that it can be difficult to estimate. The standard approach to estimating the relative risk is to fit a generalised linear model with a binomial error distribution and a log link, known as the log binomial model (60, 61). Since the log link allows predicted probabilities greater than one, convergence problems with this model are not uncommon, particularly for models containing continuous covariates or outcomes with high prevalence (60, 61). To address failed convergence with the log binomial model, several alternative approaches to relative risk estimation have been proposed, including modified Poisson regression using a log link and a robust error variance (62), and Cox regression with constant time at risk (63). For rare outcomes, where the odds ratio approximates the relative risk, another possibility is to estimate relative risks from logistic regression models (i.e. by treating the odds ratio as a relative risk). In cases where the log binomial model is deemed inappropriate due to apparent model misspecification, relative risks can also be estimated by applying marginal

or conditional standardisation to predicted probabilities obtained using logistic regression (64).

Despite the popularity of the relative risk and the widespread use of MI for handling missing data, there has been little research on the application of MI when estimating the relative risk. The primary challenge in this setting is replicating the functional form of an appropriately specified log binomial model (or equivalent, in the event of failed convergence with this model) within the imputation model. Suppose data are missing in an arbitrary pattern in outcome and exposure variables. Here the use of MI would typically entail a choice between MVNI and FCS. As described in Section 2.2.3, MVNI assumes that all variables in the imputation model follow a multivariate normal distribution, which for a binary outcome variable implies a linear relationship between the risk and other variables in the imputation model. Following imputation, continuous imputed values in the outcome need to be classified back into categories to facilitate analysis via a log binomial model. For FCS, standard software uses logistic regression to impute binary variables, which for a binary outcome assumes a linear relationship between the log odds of the risk and other variables in the imputation model. Clearly, both MVNI and FCS employ different assumptions than the intended analysis, where the log of the *risk* is assumed to be linearly related to exposure variables. It is unclear whether these differences could lead to biased or inefficient estimation.

von Hippel's MID approach could also be beneficial for relative risk estimation. Potential limitations with auxiliary variables aside, a promising feature of MID is that it may help to minimise bias introduced by a misspecified model for imputing missing outcomes (8). Should the imputation of incomplete binary outcomes using FCS or MVNI lead to biased estimation of the relative risk, this claimed strength of MID could lessen the bias. Unfortunately, little is known about the performance of MID when imputation and analysis models are incompatible, as in the current setting.

Thesis aim 2

The second aim of this thesis is to evaluate the performance of FCS and MVNI for handling missing outcome data when estimating the relative risk. Should these methods lead to biased estimates of the relative risk, a further aim is to investigate the relative merits of MID in this setting. Thesis aim 2 is addressed in Chapter 5.

3.3. Multiple imputation in randomised trials

The randomised controlled trial (RCT) is widely regarded as the gold standard design for assessing the effectiveness of health interventions. Randomisation eliminates differential selection bias by approximately balancing prognostic factors between groups, which means that a direct causal link between intervention and health outcome may be established (19). Of course, as with other study designs, the validity of causal conclusions from RCTs can be severely affected by missing outcome data.

Given the influence of evidence from RCTs on decisions concerning health policy and clinical practice, the topic of missing outcome data in RCTs has received considerable attention in the medical literature. Documents of considerable importance to biostatisticians include the International Conference on Harmonization (ICH) E9 guideline (65) and the National Research Council report on the prevention and treatment of missing data in clinical trials (1). Key recommendations in these and other guidance documents for RCTs include the need to pre-specify statistical methods for handling missing data, to state and justify the missing data mechanism assumed in the primary analysis, and to assess the robustness of findings to assumptions about the missing data mechanism in sensitivity analyses. Researchers should also detail the population parameter of interest, otherwise known as the estimand, by carefully defining both the outcome measure and the target population in which the outcome measure is defined (1, 66, 67).

This thesis focuses on the performance of MI for estimating treatment effects according to the intention to treat (ITT) principle, or equivalently, estimating the ITT estimand. For a given outcome, the ITT estimand is defined as the average effect of randomisation, irrespective of treatment received, over all randomised individuals (68). The objective of ITT is to maintain the balance in prognostic factors achieved by randomisation, which is essential for avoiding selection bias and establishing causation (69, 70). Analysis under the ITT principle is generally recommended as the preferred approach for evaluating the effectiveness of health interventions. According to the 2010 CONSORT statement, “to preserve fully the huge benefits of randomisation we should include all randomised participants in the analysis, all retained in the group to which they were allocated” (70). In a similar vein, the European Medicines Agency states that the ITT principle “is of critical importance as confirmatory clinical trials should estimate the effect of the experimental intervention in the population of patients with greatest external validity and not the effect in the unrealistic scenario where all patients receive treatment with full compliance to the treatment schedule and with a complete follow-up as per protocol” (71).

In evaluating the utility of MI for an ITT analysis, it is important to first consider whether missing outcomes should be imputed under ITT. Although some researchers have argued that imputation is necessary in order to include all randomised participants in the analysis (70, 72, 73), others have argued that an ITT analysis need only provide a valid estimate of the ITT estimand (1, 20, 74); whether or not such an analysis involves the imputation of missing outcomes is inconsequential. Given recent commentary on the importance of defining and validly estimating the estimand of interest (1), and noting that current guidance documents for RCTs do not strictly recommend imputing missing outcomes, it seems the prevailing view is that an ITT analysis need only provide a valid estimate of the ITT estimand. This is important as it means that the utility of MI must be judged solely on its ability to estimate the ITT estimand. Equivalently, statistical approaches that do not involve imputation, for example likelihood-

based methods, can be recommended over MI should they demonstrate superior statistical properties in estimating the ITT estimand.

The considerable flexibility of MI makes it an attractive option for handling missing outcome data in an ITT analysis. It is not uncommon for trialists to collect data on a large number of outcome variables. One of the key strengths of MI is its ability to handle missing data on a range of different variable types (e.g. continuous, binary, count), whether for univariate or multivariate outcomes. An added benefit of including all outcomes in a single imputation model is that observed associations between related outcomes can aid imputation. Another strength of MI is the ease with which auxiliary variables can be added to the imputation model. In RCTs, potentially useful auxiliary variables include measures of treatment compliance, proxy measures of the outcome, and even measures of the intent of participants to attend further follow-up (75). Finally, the ability of MI to be implemented under an assumption that data are MNAR makes it well suited to undertaking sensitivity analyses around a primary assumption that data are MAR (76), and as a primary method of analysis in RCTs where data are believed to be MNAR. Given the substantial flexibility of MI, it is not surprising that numerous research articles and guidance documents have advocated for its use in RCTs (e.g. (1, 12, 13, 47, 71, 77)).

Conversely, some authors have expressed a preference for the use of simpler likelihood-based approaches in RCTs (19, 24, 78). Since missing data are more likely to be restricted to the outcome in RCTs, specification of an appropriate likelihood-based method can be more straightforward than in other research settings. For missing data in a continuous multivariate outcome, likelihood-based estimation of a linear mixed model is a popular alternative to MI for estimating treatment effects under a MAR assumption (79). Although not widely known, auxiliary variables can also be incorporated into linear mixed models through joint modelling with the outcome (19, 80). For missing data restricted to a univariate outcome, the complete case likelihood is equivalent to the likelihood function of the observed data, and so a complete case analysis can be viewed as a

likelihood-based approach in this case (81). If data in the univariate outcome are MAR, the complete case analysis of this outcome is unbiased and fully efficient (18, 23, 24). Compared to MI, likelihood-based approaches offer a number of advantages: they are quicker to run, more efficient, involve fewer judgements during model-fitting, and yield a single unique estimate for a given dataset (19, 82). In addition, the issue of incompatibility between imputation and analysis models is clearly not a concern for analysis with a likelihood-based approach (82).

With the use of MI in RCTs rising dramatically in recent years (7), editors and journal reviewers are increasingly requesting to see MI used to handle missing data. For missing data restricted to a univariate outcome, there may be a reluctance to accept results from a complete case analysis given the shortcomings of this approach in general regression settings. Similarly, there is sometimes a perception that MI is the only valid option for incorporating information from auxiliary variables. However, whether it is reasonable for MI to be viewed as the gold standard approach for handling missing outcome data in RCTs is questionable. Importantly, results derived in general regression settings supporting the use of MI may not be applicable to RCTs, where missing data tend to occur primarily in the outcome and where the key exposure (randomised group) is always observed and expected to be independent of baseline covariates. With limited comparisons between MI and alternatives such as likelihood-based methods available in the literature, particularly in the estimation of treatment effects according to the ITT principle, a more rigorous investigation of the utility of MI in RCTs is needed.

Another uncertainty around the use of MI in RCTs is whether imputation should be carried out across all randomised participants or whether a separate but identical imputation model should be fitted to each randomised group. If there is interest in estimating the effect of treatment within a subgroup, the ICH E9 guideline recommends the inclusion of an interaction term between the subgroup variable and randomised group in the analysis model (65). To avoid biasing the interaction test towards the null due to incompatibility between imputation and

analysis models, the interaction term needs to be accounted for during the imputation process. Rather than specifying an interaction term within the imputation model, several authors have recommended fitting separate but identical imputation models to each randomised group (13-15). Assuming the sample size is large enough to fit separate imputation models, this strategy is appealing due to both its simplicity and its ability to facilitate subgroup analyses for any baseline covariate included in the imputation model. Often, though, subgroup analyses are not of interest, and it is unclear whether there is any merit in undertaking imputation separately by randomised group in such settings. Of particular interest is the implementation of MI in settings where interaction effects are overlooked in the analysis model in favour of producing an estimate of the average effect of treatment across subgroups.

Thesis aim 3

The third aim of this thesis is to evaluate the performance of MI for handling missing outcome data in the RCT setting and to explore the merits of imputing overall and separately by randomised group. For feasibility, the research will focus on scenarios that are commonly encountered in practice, in particular for handling missing data in a continuous or binary outcome variable measured once or repeatedly over time, and for analysis implemented under a MAR assumption. Thesis aim 3 is addressed in Chapter 6.

3.4. Multiple imputation in extended follow-up studies

Extended follow-up studies based on RCTs play an important role in assessing the longer term impacts of health interventions. Depending on the research setting, investigators may choose to initiate an extended follow-up period to learn more about disease progression, long term safety, the maintenance of early effects, or effects on longer-term, more clinically meaningful endpoints (16, 83, 84). A key benefit of initiating an extended follow-up study after the completion of an RCT is the cost saving associated with using an already established cohort. Given the

substantial investment involved in setting up a trial cohort and providing treatment, it is not surprising that many RCTs do eventually transition to extended follow-up studies (16).

Missing outcome data can pose a considerable threat to the validity of findings from extended follow-up studies. Compared to standard RCTs, the longer duration of time between randomisation and final outcome assessment in extended follow-up studies is likely to be associated with higher levels of participant attrition. In addition, investigators could choose to impose extra eligibility restrictions for inclusion into extended follow-up, for example by only recruiting participants that adhered to the protocol in the original RCT, further reducing participation rates. Depending on the information provided to participants in the original RCT, a separate consent form for entry into extended follow-up may also be required. Some participants may be unwilling to consent at this stage. Finally, participants may simply fail to provide information about a particular measure during extended follow-up. These varied sources of missing data (attrition over time, ineligibility, non-consent, and item non-response) could result in a large proportion of the original randomised cohort having missing outcome data.

An important consideration in applying MI in extended follow-up studies is whether ineligible and non-consenting participants (where applicable) should be included in the imputation model. Incorporating the full randomised cohort in the analysis preserves the benefits of randomisation, but this is likely to mean a large amount of missing data to account for and a possible mixture of missing data mechanisms at play, since reasons for missing data could differ between ineligible participants, non-consenters, and consenters. Conversely, satisfying an assumption about the missing data mechanism might be more feasible if the imputation model only included consenting participants, but then the benefits of randomisation would be diminished. In choosing a participant group to incorporate in the imputation model, important factors to consider might include the target population for the chosen estimand (e.g. all randomised for an ITT analysis) and the availability of auxiliary variables in the original RCT to aid with

the imputation of outcomes collected during extended follow-up. Whether other factors might also influence the choice of MI approach is hard to judge as there have been no published reports outlining the full scope of the missing data problem in this setting.

An informal review of published extended follow-up studies shows differences between researchers in how MI is being implemented in this setting. Two studies identified in a preliminary search of PubMed (conducted in September 2014) failed to indicate whether eligibility restrictions or separate consent processes were used (85, 86), making it difficult to understand the reasons for missing data during extended follow-up. Among studies that detailed both eligibility restrictions and separate consent processes for entry into extended follow-up, MI approaches included imputation for consenting participants in a primary analysis (87), imputation for all randomised participants in a sensitivity analysis (88), and imputation in a sensitivity analysis without any indication of the group for which results were imputed (89). In the absence of guidance documents on handling missing outcome data in extended follow-up studies, it is possible that other imputation strategies would be identified in a more thorough search of the literature.

In order to provide recommendations around the use of MI in extended follow-up studies, clearly a first step is to gain a better understanding of the missing data problem in this setting, particularly in relation to the extent and key sources of missing outcome data in this setting.

Thesis aim 4

The fourth aim of this thesis is to review the extent and common sources of missing outcome data in recently published extended follow-up studies. Based on the findings of this review, a further aim is to provide general recommendations around the implementation of MI in this setting. Thesis aim 4 is addressed in Chapter 7.

3.5. Methods for addressing thesis aims

For thesis aims 1 to 3, the performance of MI is evaluated primarily using simulation studies. In these studies, model parameters and missing data mechanisms are specified by the researcher, which means that the performance of statistical methods can be judged in relation to the known truth (90). Key statistical properties evaluated in the simulation studies include bias, measures of precision, power, and the coverage of estimated confidence intervals. For thesis aims 3 and 4, the performance of MI is also explored through application to data from the Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcome in Preterm Infants (DINO) trial (91). In DINO, n=657 preterm infants born < 33 weeks gestation were randomised between April 2001 and October 2005 to receive a high docosahexaenoic acid (DHA) or a standard DHA diet from within 5 days of commencing enteral feeds through to term. The initial DINO trial concluded following the assessment of neurodevelopmental outcomes in the children at 18 months corrected age; later an extended follow-up period was initiated to assess neurodevelopmental and growth outcomes in the children at 7 years corrected age. Ethics approval to use DINO data in this thesis was granted by the University of Adelaide Human Research Ethics Committee (approval number H-2014-239). Lastly, for thesis aim 4, the extent and common sources of missing outcome data in recently published extended follow-up studies, and statistical approaches used to handle missing outcome data in this setting, are summarised using a systematic review.

Further details on the methods for addressing the thesis aims, including descriptions of simulation parameters, additional background information on the DINO trial, and the search strategy for the systematic review, are provided in subsequent chapters.

4. Multiple imputation, then deletion

4.1. Preface

This chapter presents the first of four articles contributing to this thesis. The article, published in the *American Journal of Epidemiology*, contrasts the performance of MI and MID when auxiliary variables associated with an incomplete outcome are included in the imputation model. Previous research on MID only considered the use of auxiliary variables for efficiency gains, whereas in practice, auxiliary variables are often used to reduce bias. Another limitation of previous work is that comparisons between MI and MID only involved a small number of imputations. The purpose of this article is to provide a more comprehensive comparison between MI and MID in the presence of auxiliary variables.

4.2. Statement of authorship

Title of paper	Bias and precision of the “multiple imputation, then deletion” method for dealing with missing outcome data.
Publication status	Published
Publication details	Bias and precision of the “multiple imputation, then deletion” method for dealing with missing outcome data. <i>American Journal of Epidemiology</i> , 2015; 182(6): 528-34.


Principal author

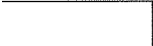
Name (Candidate)	Thomas Sullivan		
Contribution	Designed the study, simulated the data, performed all analyses, interpreted the results, drafted the manuscript and acted as corresponding author.		
Overall percentage (%)	90		
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	11/07/2017


Co-author contributions

By signing the Statement of authorship, each author certifies that:

- i. the candidate’s stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate’s stated contribution.

Name of co-author	Amy Salter		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	11/07/2017

Name of co-author	Philip Ryan		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	12/07/2017

Name of co-author	Katherine Lee		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	26/07/2017

4.3. Article

In this section, I provide the text, tables, figures, and appendices from the published manuscript.

4.3.1. Abstract

Multiple imputation (MI) is increasingly being used to handle missing data in epidemiologic research. When data on both the exposure and the outcome are missing, an alternative to standard MI is the “multiple imputation, then deletion” (MID) method, which involves deleting imputed outcomes prior to analysis. While MID has been shown to provide efficiency gains over standard MI when analysis and imputation models are the same, the performance of MID in the presence of auxiliary variables for the incomplete outcome is not well understood. Using simulated data, we evaluated the performance of standard MI and MID in regression settings where data were missing on both the outcome and the exposure and where an auxiliary variable associated with the incomplete outcome was included in the imputation model. When the auxiliary variable was unrelated to missingness in the outcome, both standard MI and MID produced negligible bias when estimating regression parameters, with standard MI being more efficient in most settings. However, when the auxiliary variable was also associated with missingness in the outcome, alarmingly MID produced markedly biased parameter estimates. On the basis of these results, we recommend that researchers use standard MI rather than MID in the presence of auxiliary variables associated with an incomplete outcome.

4.3.2. Introduction

Missing data are a widespread problem in experimental and observational research, leading to biased and inefficient parameter estimates if they are inadequately handled during the analysis. Among the more rigorous statistical approaches to handling missing data, multiple imputation (MI) has been widely

adopted due to its flexibility and relative ease of implementation (17, 29). First introduced by Rubin (4), MI uses a statistical model fitted to the observed data to estimate (impute) values for the missing data. The imputation process is repeated many times to generate multiple complete datasets, which are then analysed separately using standard statistical techniques. Finally, results from the multiple analyses are combined using Rubin's rules, which appropriately account for the uncertainty in the missing data by combining variability within and between imputed datasets (4). In its standard implementation, MI provides valid inference when data are missing at random (MAR)—that is, when the probability of missingness depends only on observed values (2).

Missing data are often evident in the outcome(s) for analysis, especially in studies involving participant follow-up. Although MI can be applied when missing data are confined to the outcome, it is most valuable when data on exposure variables are also missing (18, 92, 93). In addition to standard MI as proposed by Rubin (4), a popular method for handling missing data in outcome and exposure variables within the MI framework is von Hippel's "multiple imputation, then deletion" (MID) approach (8). As of February 11, 2015, there were 232 citations of von Hippel's article in Scopus (Elsevier B.V., Amsterdam, the Netherlands), the majority from empirical studies (e.g. (94-101)). As an illustration of the difference between standard MI and MID, consider a generalised linear model with univariate outcome Y and predictors $X = (X_1, \dots, X_p)$, where data are missing in both Y and X . Suppose also that interest lies only in estimating the parameters $\theta_{Y|X}$ that govern the conditional distribution of Y given X (e.g. regression coefficients). In both standard MI and MID, an imputation model is generated including Y and all components of X . To ensure that imputation and analysis models are consistent and to avoid biasing associations towards independence, observed and imputed values of Y are used to impute missing values for all components of X and vice versa (14, 27, 44). Following imputation, in standard MI all of the observed and imputed data for Y and X are used in the analysis of each of the completed datasets. In contrast, MID excludes (or deletes) cases with imputed Y 's from the analysis of each of the completed datasets. In other words,

analysis using standard MI involves all participants in the study, whereas analysis under MID is restricted to participants with observed outcome data. Provided that the MAR assumption is valid, the deletion of observations with imputed outcomes in MID offers two practical advantages. Firstly, for a finite number of imputations, MID has been shown to produce more precise estimates of $\theta_{Y|X}$ than standard MI (i.e. smaller standard errors, narrower confidence intervals), although efficiency gains tend to be minor unless the number of imputations is small and the proportion of missing data is high (8). Secondly, removing observations with imputed outcomes from the analysis can help to minimise the bias introduced by a misspecified model for imputing the missing outcomes (8).

The rationale behind MID is that following imputation, cases with missing outcome data do not contribute any further information about the parameters $\theta_{Y|X}$; hence, retaining these cases in the analysis only adds noise to the estimation process (8, 18). While this assertion is correct when the imputation and analysis models include the same variables (in an appropriate form), in practice these models often differ. Indeed, one of the appealing features of the MI framework is the ability to incorporate additional "auxiliary" variables into the imputation model that are not part of the substantive analysis to improve the prediction of missing values (14). In clinical trials, for example, post-randomisation measures such as treatment compliance are often used as auxiliary variables. Importantly, while both standard MI and MID benefit equally from the inclusion of auxiliary variables to improve the prediction of missing values in X , only standard MI benefits from the inclusion of auxiliary variables to predict missing values in Y (8). However, depending on the number of imputations used, the additional information provided by an auxiliary variable for Y needs to be fairly substantial for standard MI to demonstrate efficiency advantages over MID. On the basis of a simulation study involving normally distributed variables, von Hippel found that MID was more efficient than standard MI when the correlation between a single auxiliary variable and the incomplete outcome did not exceed 0.7, 0.6, and 0.5 for 2, 5, and 10 imputations, respectively (8). It is unclear whether MID would

maintain similar efficiency advantages over standard MI with a larger number of imputations.

While important, efficiency gains are not the only consideration when identifying auxiliary variables for inclusion in imputation models. Arguably the more essential role of auxiliary variables is in helping to make the MAR assumption which underlies MI more plausible. In developing high quality imputations, numerous experts have recommended the inclusion in the imputation model of auxiliary variables that are associated with the incomplete variables to be imputed, the probability of missing data, or both (e.g.(5, 6, 11, 46, 47)). It is the inclusion of auxiliary variables related to the probability of missing data that is important for satisfying the MAR assumption. Auxiliary variables related to the probability of missing data were not considered in von Hippel's original paper proposing MID (8). In a landmark study, Collins et al. (6) demonstrated via simulation that failure to incorporate information from auxiliary variables that are correlated with an incomplete outcome *and* with missingness in the outcome leads to biased inference in estimating regression coefficients from linear regression models following MI. Given the potential for auxiliary variables to reduce bias and improve efficiency, they recommended that researchers adopt inclusive strategies for selecting auxiliary variables to include in imputation models. These findings have important implications for the use of MID in studies where auxiliary information is available. Since MID is unable to take advantage of auxiliary information for an incomplete outcome, it can be argued that the approach is not entirely consistent with the inclusive strategy for variable selection when setting up an imputation model. Further, if auxiliary variables are required to satisfy a MAR assumption for the outcome, it is unclear whether including these variables in the imputation model and then deleting imputed outcomes prior to analysis could introduce bias. To our knowledge, these issues have not been investigated in the comparison of standard MI and MID.

Our aim in this paper was to evaluate the performance of standard MI and MID in regression settings where data are missing for both the outcome and the exposure

and where auxiliary variables associated with the outcome are included in the imputation model. We hypothesised that the efficiency advantages of MID would be less pronounced with a larger number of imputations, and that this approach would introduce bias in the estimation of $\theta_{Y|X}$ when the imputation model contained auxiliary variables that were additionally associated with the probability of missing data on the outcome.

4.3.3. Methods

Simulation study

We evaluated the performance of standard MI and MID in the presence of an auxiliary variable associated with an incomplete outcome by extending the earlier simulation study of von Hippel. Using the same data generation procedure, we investigated the consequences of using a larger number of imputations and allowing for missingness in the outcome to depend on an auxiliary variable.

For each simulation scenario, 1,000 complete datasets of size $n = 200$ were created. Initially, two predictor variables X_1 and X_2 were generated from a bivariate standard normal distribution with correlation ρ_{12} . An outcome Y was then produced according to the linear regression model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e$, where e was a normally distributed error term with mean 0 and variance σ^2 , and where the regression parameters $(\alpha, \beta_1, \beta_2)$ were set to (1,1,1). The proportion of the variance in Y explained by the linear regression model (R^2) was fixed by setting the variance as $\sigma^2 = 2(1 - R^2)(1 + \rho_{12})/R^2$. Next, a standard normal auxiliary variable Z was generated according to the equation $Z = \mu + \rho_{zy}Y/\text{var}(Y)$, where ρ_{zy} was the correlation between Z and Y and where μ was normally distributed with mean 0 and variance $1 - \rho_{zy}^2$. In generating complete datasets, R^2 , ρ_{12} and ρ_{zy} were independently varied. Following the simulation study of von Hippel (8), we allowed ρ_{12} and R^2 to take the values 0.2, 0.5, and 0.8, while ρ_{zy} was set to either 0.1, 0.5, or 0.9. Collectively this resulted in 27 scenarios with complete data to investigate.

Following the generation of complete datasets, values of X_2 and Y were independently set to missing according to one of two MAR mechanisms. In one setting, we replicated the "coordinated missingness" mechanism previously considered by von Hippel in which X_2 and Y were set to missing independently with probability $2p\Phi(X_1)$, where Φ is the cumulative distribution function of the standard normal distribution. Hereafter we refer to this missing data mechanism as "auxiliary independent missingness", since missingness in the outcome is conditionally independent of the auxiliary variable Z . The motivation for investigating this missing data mechanism was to evaluate the efficiency of standard MI and MID when a larger number of imputations was used; only 2, 5, and 10 imputations were considered previously. In a second setting, we considered a new missing data mechanism in which values of Y were set to missing with probability $2p\Phi([X_1 + Z]/\text{var}[X_1 + Z])$. X_2 was again set to missing with probability $2p\Phi(X_1)$. Throughout the remainder of the paper, we refer to this second missing data mechanism as "auxiliary dependent missingness". When setting values to missing, we allowed the overall proportion p of missing data in both X_2 and Y to equal 0.2 or 0.5. Together this resulted in 4 missing data patterns and 108 simulation scenarios overall.

Imputation and analysis methods

For each simulation scenario, missing values in Y and X_2 were imputed using a Markov chain Monte Carlo algorithm assuming multivariate normality (11). Y , X_1 , X_2 , and Z were all included in the imputation model. Under auxiliary independent missingness, the expected percentage of incomplete cases was 34.7% and 66.7% when the proportion of missing data in X_2 and Y was equal to 0.2 and 0.5, respectively. Based on the rule of thumb that Monte Carlo error should be acceptably small when the number of imputations equals the percentage of incomplete cases (5), the use of approximately 70 imputations is recommended for standard MI. However, since the efficiency advantages of MID are greater when the number of imputations is lower (8) and since fewer imputations are

common in practice, we chose 50 imputations as a reasonable compromise. Following imputation, the 50 complete datasets were analysed directly for standard MI and analysed following the deletion of observations with imputed outcomes for MID. Thus, for each scenario, standard MI and MID estimates were based on the same underlying imputed data. Each imputed dataset was analysed by fitting a linear regression model of the form $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e$. Of interest were the standard MI and MID estimates and 95% confidence intervals for the parameters α , β_1 and β_2 . Inference on individual parameters was obtained by combining estimates over the 50 imputed datasets using Rubin's rules (4).

Comparisons

For each simulation scenario, standard MI and MID parameter estimates across the 1,000 simulated datasets were summarised. The performance of the two approaches was assessed in terms of the bias (defined as the average difference between the parameter estimate and the true underlying value used to generate the data ($\alpha = \beta_1 = \beta_2 = 1$)) and the average estimated standard error of the parameter estimates. We also report the coverage of the estimated 95% confidence intervals, defined as the proportion of 95% confidence intervals that contained the true value. Based on 1,000 simulated datasets and a normal approximation to the binomial distribution, on 95% of occasions we would expect the coverage to lie between 0.936 and 0.964 for a nominal level of 0.95. In addition to summaries for each individual simulation scenario, mean values for the bias, average standard error, and coverage were also calculated across simulation scenarios for the two missing data mechanisms to obtain an overall measure of performance.

All statistical calculations were performed using SAS, version 9.3 (SAS Institute, Inc., Cary, North Carolina). Multiple imputation was carried out using the MI procedure, while analysis was performed using the GENMOD and MIANALYZE procedures. Starting seeds for generating variables, inducing missing data, and performing MI were varied across simulation scenarios and recorded so that results could be reproduced.

Binary variables

To investigate whether the performance of MID depends on variable type, we also performed a limited simulation study involving a binary outcome, a binary auxiliary variable, and two binary covariates. Details of this additional simulation study are outlined in the web appendix (see Section 4.3.6; also available online at the journal website).

4.3.4. Results

Table 4.1 summarises the performance of standard MI and MID under the auxiliary independent mechanism. Across the 54 simulation scenarios, both standard MI and MID exhibited negligible bias (i.e. the range of biases were consistent with Monte Carlo error), with coverage probabilities close to nominal levels throughout. In most settings, standard MI demonstrated moderate efficiency advantages over MID, with overall average standard errors (i.e. averaged across the 54 scenarios \times 1,000 datasets) for the estimated parameters α , β_1 and β_2 being at least 3% smaller with standard MI.

Table 4.1. Mean values for performance measures across 54 scenarios where missing data were induced under the auxiliary independent mechanism.

Imputation method	Parameter	Bias ^a	Range	SE	Coverage	Range
Standard MI	α	0.001	-0.016 to 0.020	0.199	0.946	0.930 to 0.962
	β_1	0.005	-0.023 to 0.042	0.259	0.946	0.926 to 0.963
	β_2	-0.006	-0.054 to 0.023	0.260	0.946	0.925 to 0.959
MID	α	0.000	-0.020 to 0.017	0.213	0.946	0.931 to 0.964
	β_1	0.004	-0.027 to 0.036	0.274	0.948	0.931 to 0.964
	β_2	-0.006	-0.056 to 0.025	0.269	0.947	0.931 to 0.966

Abbreviations: MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

^a Monte Carlo error for bias in $(\alpha, \beta_1, \beta_2) \leq (0.015, 0.025, 0.025)$ for standard MI and MID across the 54 scenarios.

The efficiency advantages of standard MI under the auxiliary independent mechanism depended most strongly on the correlation between the auxiliary variable and the outcome (ρ_{ZY}), and on the proportion of missing values in Y and X_2 (p). Table 4.2 compares the performance of the two imputation approaches for

different values of ρ_{zy} when $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5$. For $\rho_{zy} = 0.1$, the average estimated standard errors across the 1,000 imputations for the 3 parameters were approximately 1% larger using standard MI compared to MID. When the correlation ρ_{zy} was increased to 0.5, standard MI began exhibiting efficiency advantages over MID, particularly in estimating α and β_1 . In this setting, the average estimated standard errors for α and β_1 were approximately 6% smaller using MI, and they were 2% smaller for β_2 . Finally, for $\rho_{zy} = 0.9$, the average estimated standard errors were noticeably reduced with standard MI. Compared with MID, standard errors for α , β_1 , and β_2 were 28%, 27%, and 13% smaller using standard MI, respectively. A similar pattern of results was observed when the proportion of missing values in Y and X_2 was 0.2; however, absolute differences in precision were less pronounced (results not shown).

Table 4.2. Performance in scenarios where missing data were induced under the auxiliary independent mechanism for $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5^a$.

ρ_{zy}	Parameter	Standard MI			MID		
		Bias	SE	Coverage	Bias	SE	Coverage
0.1	α	-0.014	0.398	0.954	-0.013	0.393	0.958
0.1	β_1	0.018	0.401	0.957	0.019	0.396	0.960
0.1	β_2	-0.007	0.388	0.943	-0.007	0.384	0.946
0.5	α	0.020	0.371	0.948	0.017	0.393	0.946
0.5	β_1	0.012	0.375	0.954	0.009	0.398	0.955
0.5	β_2	-0.017	0.371	0.942	-0.015	0.380	0.940
0.9	α	0.003	0.283	0.951	0.001	0.392	0.956
0.9	β_1	0.008	0.290	0.951	0.005	0.395	0.957
0.9	β_2	-0.011	0.325	0.925	-0.012	0.374	0.948

Abbreviations: MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

^a Average values across the 1,000 simulations for each scenario.

As demonstrated in Table 4.3, standard MI also performed well under the auxiliary dependent mechanism. The absolute bias of standard MI was at most 0.023 across the 54 simulation scenarios for all three parameters, and the coverage probabilities remained close to nominal levels throughout. In contrast, MID showed deficiencies when the probability of missing data in the outcome variable depended on the auxiliary variable. The average bias and coverage for $(\alpha, \beta_1, \beta_2)$ across the 54 simulation scenarios was $(-0.207, -0.074, -0.017)$ and $(0.812, 0.928,$

0.947), respectively. The performance of MID suffered most when the proportion of missing data in Y and X_2 was high (0.5), when the correlation between the auxiliary variable and the outcome was high (0.9), and when the proportion of variance in Y explained by the regression model was low (0.2). Table 4.4 shows the performance of standard MI and MID under auxiliary dependent missingness for different values of ρ_{zy} when $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5$. The bias associated with MID was relatively small when $\rho_{zy} = 0.1$, although there was some evidence of undercoverage in the estimation of α and β_2 . For $\rho_{zy} = 0.5$, the bias in MID estimates was larger, particularly for α and β_1 . Finally for $\rho_{zy} = 0.9$, MID produced substantially biased estimates for α and β_1 , with coverage dropping to just 0.114 for α .

Table 4.3. Mean values for performance measures across 54 scenarios where missing data were induced under the auxiliary dependent mechanism.

Imputation method	Parameter	Bias ^a	Range	SE	Coverage	Range
Standard MI	α	-0.001	-0.014 to 0.023	0.199	0.948	0.937 to 0.962
	β_1	0.002	-0.020 to 0.017	0.255	0.947	0.932 to 0.962
	β_2	-0.004	-0.023 to 0.021	0.264	0.945	0.933 to 0.957
MID	α	-0.207	-1.329 to -0.002	0.202	0.812	0.114 to 0.961
	β_1	-0.074	-0.544 to 0.008	0.264	0.928	0.713 to 0.956
	β_2	-0.017	-0.114 to 0.012	0.271	0.947	0.932 to 0.960

Abbreviations: MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

^a Monte Carlo error for bias in $(\alpha, \beta_1, \beta_2) \leq (0.015, 0.025, 0.026)$ for standard MI and MID across the 54 scenarios.

Table 4.4. Performance in scenarios where missing data were induced under the auxiliary dependent mechanism for $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5^a$.

ρ_{zy}	Parameter	Standard MI			MID		
		Bias	SE	Coverage	Bias	SE	Coverage
0.1	α	-0.004	0.400	0.945	-0.137	0.360	0.935
0.1	β_1	-0.020	0.376	0.943	-0.076	0.365	0.941
0.1	β_2	-0.006	0.399	0.937	-0.008	0.395	0.934
0.5	α	0.001	0.374	0.957	-0.622	0.360	0.585
0.5	β_1	-0.009	0.357	0.952	-0.277	0.362	0.883
0.5	β_2	0.001	0.384	0.945	-0.026	0.385	0.935
0.9	α	-0.012	0.283	0.954	-1.111	0.350	0.114
0.9	β_1	-0.001	0.282	0.948	-0.478	0.346	0.713
0.9	β_2	-0.023	0.330	0.949	-0.114	0.362	0.955

Abbreviations: MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

^a Average values across the 1,000 simulations for each scenario.

To more accurately demonstrate the bias introduced by MID in the presence of an auxiliary variable associated with the outcome and with missingness in the outcome, we performed additional simulations for $\rho_{12} = 0.2$ and $R^2 = 0.2$, where we varied the correlation between the auxiliary variable and the outcome (ρ_{zy}) in increments of 0.1. The performance of standard MI and MID in estimating α and β_1 are plotted in Figure 4.1. As shown in Figure 4.1A, estimates of β_1 were close to the true value for both standard MI and MID when the proportion of missing data in Y and X_2 was 0.2. However, when the proportion of missing data in Y and X_2 was increased to 0.5, MID exhibited bias, even for small values of ρ_{zy} , with the magnitude of the bias increasing linearly with the correlation ρ_{zy} . A similar pattern of results was observed for α (Figure 4.1B), although for this parameter MID also exhibited some bias when the proportion of missing data in Y and X_2 was 0.2.

In line with results for continuous outcomes, standard MI performed well when missing data in a binary outcome depended on an auxiliary variable, but coefficient estimates in a logistic regression model were biased with MID (see web appendix, Table 4.5). Once again the magnitude of the bias of MID depended on the strength of the association between the outcome and the auxiliary variable.

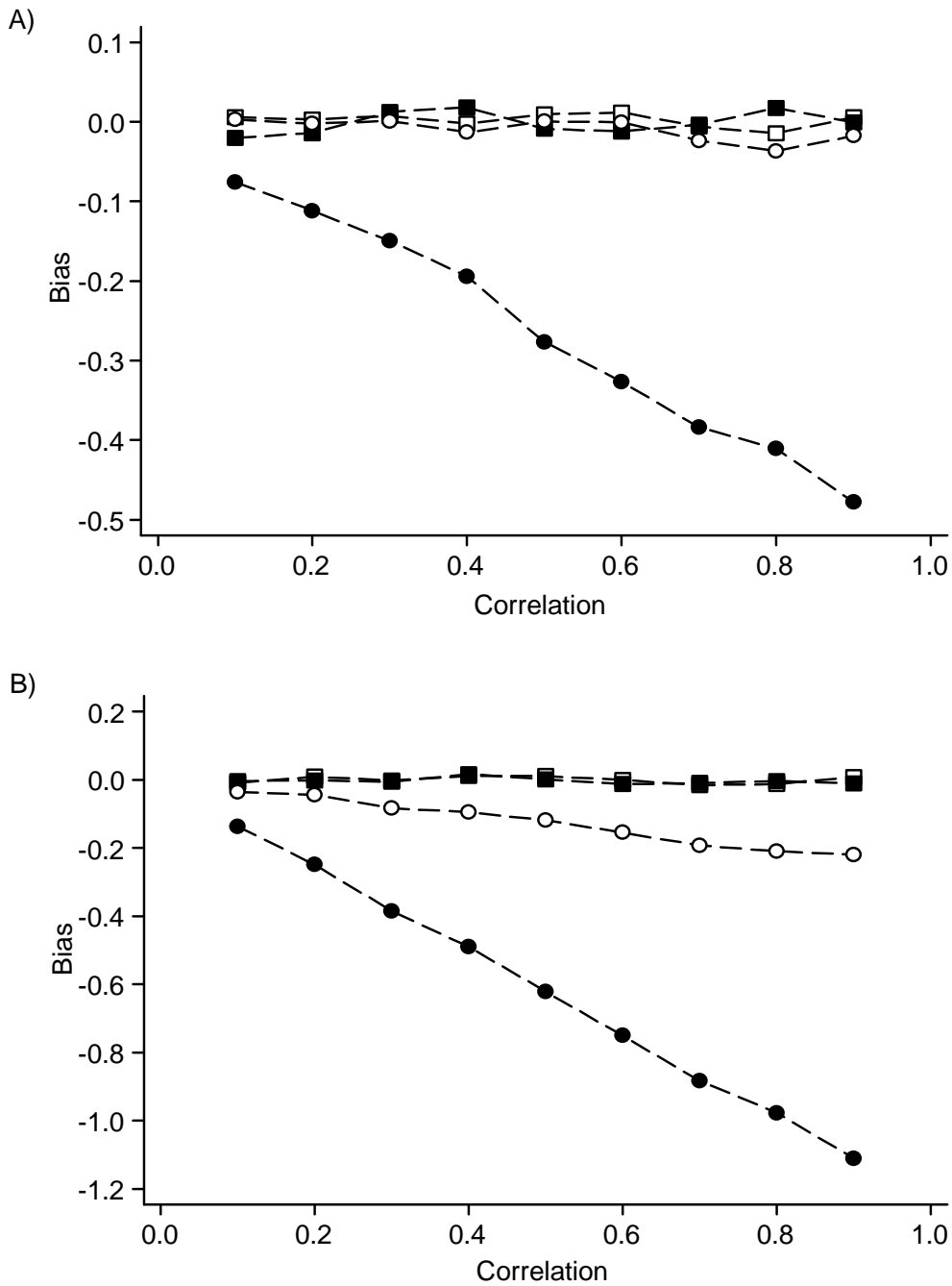


Figure 4.1. Bias under the auxiliary dependent mechanism in the estimation of β_1 (A) and α (B) for $\rho_{12} = 0.2$ and $R^2 = 0.2$. Correlation on the x-axis represents the correlation between the outcome Y and the auxiliary variable Z . Results are for multiple imputation with $p = 0.2$ (white squares), multiple imputation with $p = 0.5$ (black squares), “multiple imputation, then deletion” (MID) with $p = 0.2$ (white circles), and MID with $p = 0.5$ (black circles).

4.3.5. Discussion

In this study, we evaluated the performance of standard MI and MID when the imputation model was enriched by auxiliary information for the incomplete outcome. In line with previous results, both standard MI and MID exhibited negligible bias in estimating regression parameters when an auxiliary variable associated with the incomplete outcome, but not with missingness in the outcome, was added to the imputation model. We have now demonstrated that when the auxiliary variable is also related to missingness in the outcome and hence is required in the imputation model to satisfy the MAR assumption, MID produces biased estimates of regression parameters, whereas standard MI does not. These results have important implications for the use of MID in applied research.

When the auxiliary variable was unrelated to missingness in the outcome, results demonstrated that the precision of MID was only marginally better than that of standard MI for a weak correlation between the auxiliary variable and the outcome. Conversely, standard MI was noticeably more efficient for moderate-to-strong correlations between the auxiliary variable and the outcome. The results are in line with those observed previously for 10 or fewer imputations (8), however, in our study, the efficiency advantages of standard MI were greater with 50 imputations. This suggests that the intended number of imputations is an important factor to take into account when choosing between standard MI and MID based solely on efficiency considerations. Although early texts on MI suggested that 10 or fewer imputations are usually adequate (25, 27, 36), more recent recommendations state that the number of imputations should be much larger (i.e. 20 to 100) (5, 45). Since increasing the number of imputations entails greater precision, standard MI with a large number of imputations should be preferred over MID if the primary goal is to maximise efficiency. In light of continuing improvements in computational power and analytical software, standard MI with a large number of imputations should be feasible in most practical settings.

When missingness in the outcome depended on the auxiliary variable, MID produced biased estimates of regression parameters, with the magnitude of the bias being positively associated with the amount of missing data and the correlation between the auxiliary variable and the outcome. Effectively, MID discarded the information about the outcome provided by the auxiliary variable, leading to the violation of a MAR assumption that was otherwise satisfied under standard MI. The results suggest that MID is not an optimal strategy in the presence of auxiliary variables that are associated with missingness in the outcome. In our view, failing to exploit the information offered by auxiliary variables and potentially introducing serious bias into the analysis for small potential gains (or possible losses) in precision is a poor trade-off. This leaves researchers with two choices for implementing MI when auxiliary information for an incomplete outcome is available: 1) imputing using a model that excludes auxiliary variables associated with the incomplete outcome and proceeding with MID or 2) incorporating these auxiliary variables into the imputation model and employing a standard MI analysis. Given the potential value of auxiliary variables for bias reduction and efficiency gains, we believe the latter option is preferable in most settings.

Clearly, results based on a restricted simulation study such as this cannot be generalised to all applied settings. For example, in this study we did not consider scenarios with missingness in auxiliary variables, multiple auxiliary variables, or more complex regression models, all of which are common in practice. Further, in all simulation scenarios the association between the auxiliary variable and the probability of missing data in the outcome was fixed; previous research has shown that the strength of this association is an important determinant of the bias associated with failing to include an auxiliary variable in the imputation model (48). While the simulation study illustrates the potential for introducing bias using MID, the extent of this bias will depend on specific characteristics of the individual study. Associations involving auxiliary variables may be weaker than those considered in this study, and hence the bias introduced by MID may not be of practical importance in many settings (6, 48, 49). Alternatively, researchers

may have access to a large number of auxiliary variables, which collectively could have a dramatic influence on bias and efficiency. Thus, while the bias in estimating regression coefficients with 20% missing data was moderate in the current study, it could be larger in other settings with similar amounts of missing data.

A further limitation of this study is that it only considered a MAR mechanism and a correctly specified imputation model. Both conditions may not be met in practice. Although a MAR assumption is often plausible, data may instead be missing not at random, which occurs when the probability of missingness depends on unobserved values (2). Unless missingness occurs by design, it is impossible to tell whether data are truly MAR or missing not at random based only on observed values. If imputation is performed under a MAR assumption when data are in fact missing not at random, in general this will lead to biased inference, although auxiliary variables can help to mitigate this bias (6). Since MID is unable to incorporate information about an incomplete outcome from auxiliary variables, it may be that this approach would produce more biased estimates than standard MI when data are missing not at random, although this remains to be investigated. In choosing between standard MI and MID, another important consideration is the ability to adequately specify the imputation model. One argument for using MID is that removing imputed outcomes from the analysis will reduce the bias introduced by a misspecified model for imputing outcomes. Whether this is important in practice is unclear. Popular methods of imputation such as multivariate normal imputation and fully conditional specification are known to be fairly robust to model misspecification (e.g.(5, 11, 40, 43)), while ad hoc approaches such as predictive mean matching can be used when there is uncertainty surrounding relationships between variables in the imputation model (5, 41). Thus, even in settings where there is considerable uncertainty in specifying an appropriate imputation model, we would still recommend proceeding with standard MI when auxiliary information for an incomplete outcome is available.

In summary, MID can lead to biased estimation when auxiliary variables that are associated with missingness in an incomplete outcome are included in the imputation model. Once a decision has been made to include auxiliary variables in the imputation model, whether to satisfy a MAR assumption or to improve precision, we recommend retaining this information in the analysis and using a standard MI approach.

4.3.6. Web appendix

Methods

For each simulation scenario, 1,000 complete datasets of size $n = 500$ were created. A larger sample size was considered for binary outcomes to reduce the likelihood of observing zero-cells in cross-tabulations involving the outcome. Initially, two dependent binary variables X_1 and X_2 were generated with success probability 0.5 and with an odds ratio for their association (i.e. $\text{odds}[X_2 = 1|X_1 = 1]/\text{odds}[X_2 = 1|X_1 = 0]$) of 2.25. A binary outcome Y was then generated according to the logistic regression model $\text{logit } P(Y = 1) = \alpha + \beta_1 X_1 + \beta_2 X_2$, where the regression parameters $(\alpha, \beta_1, \beta_2)$ were set to $(-1, 1, 1)$. Next, a binary auxiliary variable Z was generated according to the equation $P(Z = 1) = \lambda + \tau Y$, with values of λ and τ chosen to give Z an overall success probability of 0.5 and an odds ratio for the association with Y (i.e. $\text{odds}[Z = 1|Y = 1]/\text{odds}[Z = 1|Y = 0]$) of either 2, 5, or 10 (three scenarios). Following the generation of complete datasets, values of X_2 were set to missing with probability $0.2 + 0.6X_1$. Independently values of Y were set to missing with probability $0.2 + 0.3X_1 + 0.3Z$ (i.e. missing data in Y depended on the auxiliary variable Z). The missing data mechanism resulted in 50% missing data for both X_2 and Y .

For each of the three simulation scenarios, missing values in Y and X_2 were imputed using fully conditional specification (9, 10) with 50 cycles and 50 imputations. Y , X_1 , X_2 and Z were specified as binary variables in the imputation model, with Y and X_2 imputed using logistic regression models. Following

imputation, the 50 complete datasets were analysed directly for standard multiple imputation (MI) and analysed following the deletion of observations with imputed outcomes for multiple imputation, then deletion (MID). Each imputed dataset was analysed by fitting a logistic regression model of the form $\text{logit } P(Y = 1) = \alpha + \beta_1 X_1 + \beta_2 X_2$. Of interest were the standard MI and MID estimates and 95% confidence intervals for the parameters α , β_1 , and β_2 . Inference on individual parameters was obtained by combining estimates over the 50 imputed datasets using Rubin's rules (4). Performance across the 1,000 simulated datasets for each parameter was summarised using the bias, average estimated standard error and coverage. All statistical calculations were performed using SAS version 9.3 (SAS Institute, Inc., Cary, North Carolina).

Results

Table 4.5 summarises the performance of standard MI and MID for the 3 simulation scenarios for binary outcomes. In line with results for continuous outcomes, standard MI performed well when missing data in the binary outcome depended on the auxiliary variable. Bias was negligible for all parameters across all scenarios, and coverage probabilities remained close to nominal levels throughout. In contrast, MID produced biased parameter estimates of α and β_1 , with the magnitude of bias increasing with the strength of the association between the auxiliary variable and the outcome. Coverage probabilities and standard errors for these parameters also suffered with MID. Of note, MID exhibited negligible bias in estimating β_2 . This finding is not unexpected given the symmetrical properties of the odds ratio and the missing data mechanism considered. Since the probability of missing data in our example depended on X_1 and Y (via the auxiliary variable Z) but not X_2 , an analysis restricted to cases with complete data would be expected to provide an unbiased estimate of β_2 (102). In terms of precision, average standard errors for β_2 were larger with MID than with standard MI, as seen for the estimation of α and β_1 .

Table 4.5. Performance in scenarios for a binary outcome^a.

OR_{zy}	Parameter	Standard MI			MID		
		Bias ^b	SE	Coverage	Bias ^b	SE	Coverage
2	α	-0.012	0.242	0.955	-0.094	0.242	0.944
2	β_1	0.013	0.304	0.946	-0.055	0.305	0.942
2	β_2	-0.004	0.362	0.955	-0.002	0.364	0.957
5	α	-0.018	0.241	0.954	-0.199	0.246	0.908
5	β_1	0.017	0.294	0.956	-0.130	0.306	0.922
5	β_2	0.005	0.357	0.951	0.007	0.367	0.957
10	α	-0.007	0.235	0.949	-0.252	0.246	0.868
10	β_1	0.005	0.283	0.948	-0.208	0.307	0.894
10	β_2	0.009	0.346	0.955	0.011	0.365	0.956

Abbreviations: OR_{zy} , odds ratio for the association between Z and Y; MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

^a Average values across the 1,000 simulations for each scenario.

^b Monte Carlo error for bias in $(\alpha, \beta_1, \beta_2) \leq (0.008, 0.010, 0.011)$ for standard MI and MID across the 3 scenarios.

*** End of published article ***

4.4. Additional discussion

Another possible method for handling missing outcome data within the MI framework that was not mentioned in the published article is to delete observations with missing outcomes prior to fitting the imputation model, an approach termed “deletion, then multiple imputation” (DMI) (8). In the absence of auxiliary variables, von Hippel found that DMI was marginally more biased and less efficient than MID across a range of simulation scenarios where data were MAR in the outcome and exposure variables. However, in settings where participants with missing outcome data tended to have complete data on exposure variables, and vice versa, DMI performed considerably worse than MID in terms of bias and precision (8). Based on these results, von Hippel discouraged the use of DMI in practice, and subsequently this approach was not considered when the article in this chapter was conceived. More recently, Kontopantelis et al. (103) evaluated imputation strategies for handling missing outcome data and observed little difference in performance between DMI, MID, and MI, both in the absence of auxiliary variables and in settings where a single auxiliary variable was used for efficiency gains (but not bias reduction). Based on these results, the authors concluded that the choice of imputation approach makes little difference in

practice; the important thing is that the outcome is included in the imputation model. In light of this recent recommendation, an additional investigation into the performance of DMI in the presence of auxiliary variables for the outcome seems warranted.

The statistical properties of DMI were evaluated via simulation using the data generation procedures and missing data mechanisms from the main article (see Section 4.3.3). Results obtained for DMI were then compared with findings for MID and standard MI. Under the auxiliary independent mechanism, DMI produced unbiased parameter estimates that were slightly less efficient than corresponding MID estimates (which were also unbiased). Compared to MID, average estimated standard errors for α , β_1 , and β_2 were approximately 2.1%, 1.0%, and 0.4% larger with DMI, respectively. However, like MID, DMI was at times substantially less precise than standard MI. As illustrated in Figure 4.2 for the parameter β_1 , where $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5$, standard MI exhibited noticeable efficiency advantages over DMI (and MID) as the correlation between the outcome and the auxiliary variable (ρ_{zy}) increased to 0.9. Interestingly, average estimated standard errors for DMI and MID appeared invariant to ρ_{zy} , suggesting that these approaches were not incorporating any of the information provided by the auxiliary variable.

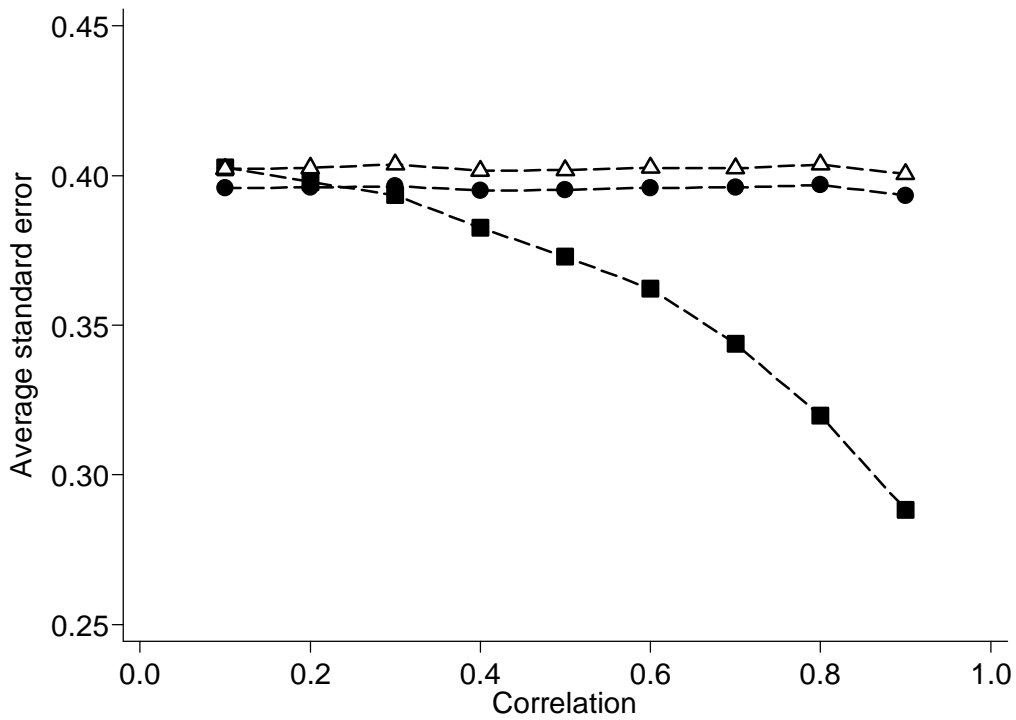


Figure 4.2. Average estimated standard errors for β_1 under the auxiliary independent mechanism where $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5$. Correlation on the x-axis represents the correlation between the outcome Y and the auxiliary variable Z . Results are for MI (black squares), MID (black circles), and DMI (white triangles).

Similarities in performance between DMI and MID also extended to settings where the probability of missing data in the outcome depended on the auxiliary variable, albeit with DMI marginally more biased and less efficient than MID. Importantly, both DMI and MID were inferior to standard MI. As evident in Figure 4.3 for the parameter β_1 , where $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5$, the bias of both DMI and MID became progressively more pronounced as ρ_{zy} increased, while standard MI remained unbiased. A similar pattern of results was observed for $p = 0.2$, although absolute differences in bias were less pronounced (results not shown).

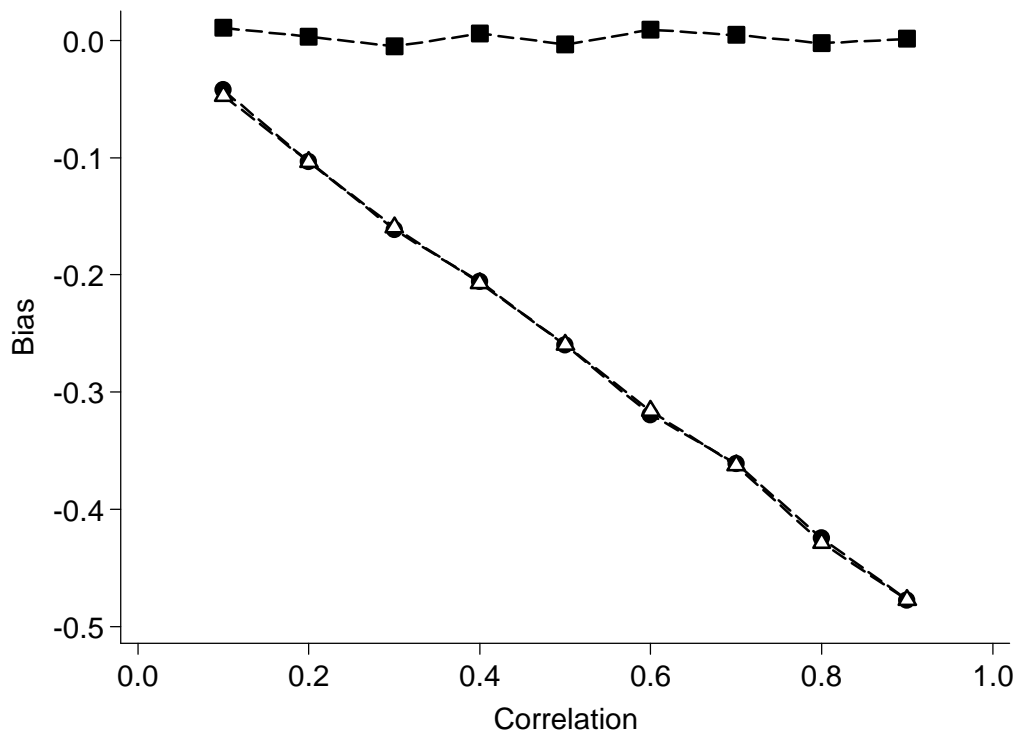


Figure 4.3. Bias for β_1 under the auxiliary dependent mechanism where $\rho_{12} = 0.2$, $R^2 = 0.2$, and $p = 0.5$. Correlation on the x-axis represents the correlation between the outcome Y and the auxiliary variable Z . Results are for MI (black squares), MID (black circles), and DMI (white triangles).

Collectively, the results from this additional simulation study do not alter the main message from the published article, which is that it is preferable to employ standard MI when the imputation model contains auxiliary variables for the incomplete outcome.

5. Multiple imputation for estimating the relative risk

5.1. Preface

This chapter presents the second article contributing to this thesis, published in *BMC Medical Research Methodology*. The primary aim of the article is to evaluate the performance of standard model-based methods of MI for handling missing outcome data when estimating the relative risk. Given the potential for bias due to a misspecified imputation model, a further aim is to investigate whether removing imputed outcome values using MID improves estimation. Given the findings of Chapter 4, any potential benefits of MID for relative risk estimation should be weighed against the limitations of this approach in the presence of auxiliary variables.

5.2. Statement of authorship

Title of paper	Multiple imputation for handling missing outcome data when estimating the relative risk.
Publication status	Published
Publication details	Multiple imputation for handling missing outcome data when estimating the relative risk. <i>BMC Medical Research Methodology</i> , 2017; 17(1): 134.

Principal author

Name (Candidate)	Thomas Sullivan		
Contribution	Designed the study, simulated the data, performed all analyses, interpreted the results, drafted the manuscript and acted as corresponding author.		
Overall percentage (%)	90		
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	11/07/2017

Co-author contributions

By signing the Statement of authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of co-author	Katherine Lee		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	26/07/2017

Name of co-author	Philip Ryan		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	12/07/2017

Name of co-author	Amy Salter		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	11/07/2017

5.3. Article

In this section, I provide the text, tables, and appendices from the revised version of the manuscript.

5.3.1. Abstract

Background: Multiple imputation is a popular approach to handling missing data in medical research, yet little is known about its applicability for estimating the relative risk. Standard methods for imputing incomplete binary outcomes involve logistic regression or an assumption of multivariate normality, whereas relative risks are typically estimated using log binomial models. It is unclear whether misspecification of the imputation model in this setting could lead to biased parameter estimates.

Methods: Using simulated data, we evaluated the performance of multiple imputation for handling missing data prior to estimating adjusted relative risks from a correctly specified multivariable log binomial model. We considered an arbitrary pattern of missing data in both outcome and exposure variables, with missing data induced under missing at random mechanisms. Focusing on standard model-based methods of multiple imputation, missing data were imputed using multivariate normal imputation or fully conditional specification with a logistic imputation model for the outcome.

Results: Multivariate normal imputation performed poorly in the simulation study, consistently producing estimates of the relative risk that were biased towards the null. Despite outperforming multivariate normal imputation, fully conditional specification also produced somewhat biased estimates, with greater bias observed for higher outcome prevalences and larger relative risks. Deleting imputed outcomes from analysis datasets did not improve the performance of fully conditional specification.

Conclusions: Both multivariate normal imputation and fully conditional specification produced biased estimates of the relative risk, presumably since both use a misspecified imputation model. Based on simulation results, we recommend researchers use fully conditional specification rather than multivariate normal imputation and retain imputed outcomes in the analysis when estimating relative risks. However fully conditional specification is not without its shortcomings, and so further research is needed to identify optimal approaches for relative risk estimation within the multiple imputation framework.

5.3.2. Introduction

The relative risk is a summary measure of effect for binary outcomes that is often of interest in medical research (54-57). Unlike the odds ratio, the relative risk is simple to interpret and collapsible across covariate strata (58). For rare outcomes, relative risks may be estimated from logistic regression models, since the odds ratio approximates the relative risk in this case (57). For more common outcomes, the odds ratio overestimates the relative risk and so alternatives to logistic regression are required to estimate the relative risk. A standard approach to estimating the relative risk directly is to fit a generalised linear model with a binomial error distribution and a log link, known as the log binomial model (60, 61). Since the log link allows predicted probabilities greater than one, convergence problems with this model are not uncommon, particularly for models containing continuous covariates or outcomes with high prevalence (60, 61). Several alternative approaches to relative risk estimation have been proposed to address failed convergence with the log binomial model, with modified Poisson regression using a log link and a robust error variance (62) one of the more commonly used methods.

A common feature of epidemiologic investigations is the occurrence of missing data, which can result in biased and inefficient parameter estimates if inadequately handled during the statistical analysis. Among the more rigorous approaches to handling missing data, multiple imputation (MI) (4) has been widely adopted due

to its flexibility and availability in statistical software packages (7). MI involves fitting a statistical model to the observed data to estimate values for the missing data. To incorporate missing data uncertainty, multiple values are imputed for each missing observation, producing multiple complete datasets. Following analysis, parameter estimates from the multiple datasets are appropriately combined to give a single MI estimate. Standard implementations of MI assume that data are missing at random (MAR), which occurs when the probability of missing data depends only on observed data (2). Provided this assumption is met and statistical models used for imputation and analysis are correctly specified, MI produces consistent and asymptotically efficient parameter estimates (4).

For arbitrary patterns of missing data (i.e. missing data occurring in any variable, in any pattern across variables), the two standard model-based methods of MI are fully conditional specification (FCS) (5, 9, 10), also known as chained equations, and multivariate normal imputation (MVNI) (11). FCS involves specifying a series of univariate imputation models, one for each variable with missing data. Standard software uses logistic regression to impute incomplete binary outcomes, which assumes a linear relationship between the log odds of the risk and other variables in the imputation model. Incomplete covariates can similarly be imputed using appropriate univariate models (e.g. linear regression for continuous covariates). In contrast, MVNI assumes that all variables in the imputation model follow a multivariate normal distribution. For incomplete binary outcomes, an additional rounding step is also required following MVNI to convert continuous imputed values to binary values suitable for analysis (40). Although FCS and MVNI have been evaluated in settings where the goal is to estimate the odds ratio using logistic regression (9, 40, 43), little is known about their performance when the aim is to estimate the relative risk. Importantly, it is unclear whether imputing outcomes using logistic regression in FCS or under a multivariate normal assumption in MVNI could lead to biased or inefficient estimation when the analysis involves a log binomial model.

A popular alternative to the standard implementation of MI for handling missing data in both outcome and exposure variables is the “multiple imputation, then deletion” approach (MID), where observations with imputed outcomes are excluded from the analysis (8). Although MID is not advisable when the imputation model contains auxiliary variables for the outcome (i.e. variables that are not part of the analysis but which help to predict missing outcome values) (104), the approach can offer small efficiency gains over standard MI when imputation and analysis models are the same. Of relevance to the estimation of relative risks, it has been argued that removing imputed outcomes prior to analysis can help to minimise the bias introduced by a misspecified imputation model for the outcome (8). Should the imputation of incomplete binary outcomes using FCS or MVNI lead to biased estimation of the relative risk, this claimed strength of MID could lessen this bias.

This article aims to (i) evaluate the performance of FCS and MVNI for handling missing outcome data when estimating the relative risk, and (ii) investigate whether deleting imputed outcomes prior to analysis improves the performance of FCS and MVNI in this setting. The rest of the article is set out as follows. In the next section, we describe the methods of FCS and MVNI in more detail, drawing attention to potential limitations. This is followed by an outline of the simulation methods used to address the article aims, and a summary of the simulation results. Finally, we conclude the article by discussing key findings and providing recommendations for practice.

5.3.3. Methods

Fully conditional specification

FCS involves specifying a series of univariate imputation models, one for each variable with missing data (5, 9, 10), with models tailored according to the distribution of the variable being imputed. For each variable with missing data, the FCS algorithm begins by replacing missing values with randomly selected

observed values or the mean value for the same variable. Imputations are then generated by estimating each univariate model in turn, restricted to participants with observed values for the variable being considered and using imputed values for other variables; at each stage missing values are replaced by draws from their posterior predictive distribution. This process continues until all incomplete variables have been imputed and is repeated several times in order to stabilise the results, leading to the generation of a single imputed dataset. Additional imputed datasets are obtained by independently repeating this process.

Despite its flexibility, FCS is not without limitations. One concern with the approach is the possibility of specifying univariate imputation models where the conditional distributions implied do not correspond to a valid joint distribution. A potential consequence of this is that results could vary according to the ordering of univariate imputation models within the FCS procedure. Fortunately this issue appears to have little impact on results in practice (9, 10, 41, 43). Another drawback of FCS is that it can be time consuming to implement in settings containing a large number of incomplete variables, since univariate imputation models need to be specified for each incomplete variable in the imputation model.

Multivariate normal imputation

MVNI is a joint modelling approach to imputation where all variables in the imputation model are assumed to follow a multivariate normal distribution. First implemented by Schafer (11), MVNI uses a Markov chain Monte Carlo algorithm (known as data augmentation) for imputation. Initially, missing values are imputed based on assumed starting parameter values for the multivariate normal distribution. These are typically obtained from available data using the expectation-maximisation algorithm. Next, updated parameter values for the multivariate normal distribution are drawn from their posterior distribution based on the observed and imputed data. This iterative process of imputing missing values and drawing updated parameter values continues until these values converge to a stationary distribution (11, 36). Following these “burn-in”

iterations, a set of imputed values is taken. In order to reduce dependence between imputations, additional iterations are performed before the next set of imputed values is obtained.

Due to its strong theoretical underpinnings, MVNI is an appealing method when multivariate normality holds, but such an assumption is not always realistic, particularly when the imputation model contains binary variables. Although several authors have reported good performance with MVNI for binary variables (11, 37, 40, 41), it remains difficult to make global statements about the robustness of this approach to violations of multivariate normality, whether in the specific case of binary variables or more generally.

Simulation study

The performance of FCS and MVNI for handling missing outcome data when estimating the relative risk was evaluated using data simulation. In order to attribute any deficiencies in performance to the method of MI, rather than getting caught up in complexities of the data, we focused on relatively simple simulation scenarios.

In each simulation scenario, 2000 datasets of size $n = 1000$ were generated from the log binomial model $\log P(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, where X_1 and X_2 were binary or normally distributed exposure variables and Y was the binary outcome. A relatively large sample size was chosen to avoid zero cells in cross-tabulations involving the outcome. Following generation of complete datasets, values in X_2 and Y were set to missing according to a specified MAR mechanism to produce an arbitrary pattern of missing data in these two variables. Missing values were then multiply imputed using FCS or MVNI with $m = 20$ imputations. For FCS, missing values in Y were imputed using a logistic regression model, while imputations for binary or normally distributed X_2 were generated from a logistic or linear regression model respectively. A total of 20 cycles were used for each imputation, with the outcome imputed last. For MVNI,

missing values were imputed using a Markov chain Monte Carlo algorithm with a burn-in of 200 iterations. Following imputation with MVNI, imputed values in the outcome were rounded to binary values using adaptive rounding, which has been recommended over alternative rounding techniques (40). Finally, complete datasets either retaining or deleting imputed outcomes were analysed using log binomial models (or modified Poisson regression as appropriate), with parameter estimates for β_1 and β_2 combined across datasets using Rubin's rules (4). Since the outcome Y was generated under the analysis model, any deficiencies in performance could be attributed to the method of MI. For reference, a complete case analysis (CCA) restricted to participants with complete data on both Y and X_2 was also performed in each simulation scenario.

Simulation study 1: categorical exposures

In simulation study 1, X_1 and X_2 were generated as binary variables with a prevalence of 0.50 and a relative risk for their association ($RR(X_1, X_2)$) of 2 or 3, to induce moderate or strong confounding respectively. In simulating values for the outcome Y , β_1 and β_2 were both set to $\log(2)$ or $\log(3)$ to give conditional relative risks (i.e. $RR(Y, X_1|X_2)$ and $RR(Y, X_2|X_1)$) of 2 or 3. Lastly the intercept β_0 was chosen to give an overall outcome prevalence of 0.10 or 0.30. Following generation of complete datasets, values in Y and X_2 were set to missing according to one of two MAR mechanisms:

- 1) Coordinated: $\text{logit } P(Y \text{ missing}) = \text{logit } P(X_2 \text{ missing}) = \alpha + \lambda X_1$.
- 2) Opposite: $\text{logit } P(Y \text{ missing}) = \alpha + \lambda X_1$, $\text{logit } P(X_2 \text{ missing}) = \alpha + \lambda(1 - X_1)$.

Under the coordinated mechanism, participants with missing data were often missing both Y and X_2 , whereas under the opposite mechanism, participants with missing data tended to be missing either Y or X_2 (but not both). For both mechanisms, the parameter λ was set to 1 or 2 to indicate a moderate or strong missing data mechanism respectively, while α was chosen to produce 30%

missing data in Y and X_2 . Collectively this resulted in 4 missing data patterns and 32 simulation scenarios. Following imputation, complete datasets were analysed using log binomial models. Provided MVNI was applied with adaptive rounding for imputed values in X_2 (in addition to Y), there were no convergence issues with the log binomial model in this setting.

Simulation study 2: continuous exposures

For simulation study 2, X_1 and X_2 were generated from a bivariate normal distribution with mean 0, variance 0.20 and correlation ($\text{corr}(X_1, X_2)$) 0.30 or 0.70. Again β_1 and β_2 were set to $\log(2)$ or $\log(3)$ to give conditional relative risks of 2 or 3, while β_0 was chosen to give an outcome prevalence of 0.10 or 0.30. One concern when simulating data under a log binomial model with unbounded continuous covariates is the possibility of generating ‘success’ probabilities greater than one. In choosing the variance for X_1 and X_2 , we sought to maximise the size of standardised conditional relative risks while minimising the occurrence of invalid success probabilities. With a variance of 0.20, invalid success probabilities were rare, except in settings involving an outcome prevalence of 0.30 and conditional relative risks of 3 (where 5.4% of success probabilities exceeded one). Following previous simulation studies exploring the relative risk (e.g. (105)), X_1 and X_2 were resampled in these instances to ensure valid success probabilities.

Letting $Z_1 = X_1/\sqrt{\text{var}(X_1)}$, the coordinated and opposite missing data mechanisms were adapted for the continuous setting as follows:

- 1) Coordinated: $\text{logit } P(Y \text{ missing}) = \text{logit } P(X_2 \text{ missing}) = \alpha + \lambda Z_1$.
- 2) Opposite: $\text{logit } P(Y \text{ missing}) = \alpha + \lambda Z_1$, $\text{logit } P(X_2 \text{ missing}) = \alpha - \lambda Z_1$.

In line with simulation study 1, λ was set to 1 or 2 and α was chosen to produce 30% missing data in Y and X_2 . Again this resulted in 4 missing data patterns and 32 simulation scenarios. As non-convergence with the log binomial model was a

considerable problem in this setting, often occurring for some but not all imputed datasets within a single simulation, we elected to analyse all complete datasets using modified Poisson regression.

Comparisons

The performance of the MI approaches in estimating parameters β_1 and β_2 was evaluated in terms of bias (average difference between estimate and true value) and the coverage of estimated 95% confidence intervals (proportion of 95% confidence intervals containing the true value). With 2000 simulated datasets per simulation scenario, on 95% of occasions the coverage is expected to lie between 0.94 and 0.96 for a true coverage of 0.95. For each parameter, the average within-simulation estimated standard error (denoted the average standard error), the standard error of parameter estimates across simulated datasets (denoted the empirical standard error), and the mean square error (average squared difference between the estimate and the true value) were also derived. All analyses were performed in SAS version 9.4 (SAS Institute, Inc., Cary, North Carolina). Multiple imputation was carried out using the MI procedure, while analysis was performed using the GENMOD and MIANALYZE procedures. The SAS code for implementing the simulation study is available in web appendix A (see Section 5.3.6).

5.3.4. Results

Simulation study 1: categorical exposures

Table 5.1 displays results for the categorical exposure setting in scenarios with a strong missing data mechanism ($\lambda = 2$), where $RR(X_1, X_2) = 2$ and $\beta_1 = \beta_2 = \log(3)$. Similar results were observed for $RR(X_1, X_2) = 3$, while absolute biases of the imputation approaches were smaller in magnitude when $\lambda = 1$ and $\beta_1 = \beta_2 = \log(2)$. Full results for all simulation scenarios are available in web appendix B (see Section 5.3.6). MVNI performed poorly across the 32 simulation

scenarios, consistently producing estimates of β_2 that were biased towards the null (bias range -0.32 to -0.10). The bias of -0.32 shown in Table 5.1 for an outcome prevalence of 0.30 under the coordinated mechanism equates to a relative risk estimate of 2.19 compared with the true value of 3; coverage was just 0.55 in this scenario. Bias was less of a concern for β_1 (bias range -0.08 to 0.07). Deleting imputed outcomes following MVNI led to some reduction in absolute bias for β_2 , although estimates for β_1 were moderately biased away from the null with this approach (bias range 0.02 to 0.11). Interestingly, average and empirical standard errors were noticeably increased by the deletion of imputed outcomes following MVNI. Compared to MVNI (without deletion), MVNI with deletion led to small increases in the mean square error for β_1 , but tended to decrease the mean square error for β_2 .

In contrast to MVNI, FCS performed fairly well for categorical exposures, with absolute bias only exceeding 0.10 for the coefficient β_2 in scenarios involving a strong coordinated mechanism, an outcome prevalence of 0.30 and where $\beta_1 = \beta_2 = \log(3)$. Excluding simulation scenarios where the bias for β_2 exceeded 0.10, the coverage of estimated 95% confidence intervals for β_1 and β_2 remained close to nominal levels (range 0.93 to 0.96). Compared to FCS (without deletion), FCS with deletion led to small reductions in absolute bias for β_2 under the coordinated mechanism for an outcome prevalence of 0.30, but slight increases in absolute bias under the opposite mechanism for the same outcome prevalence. There was little difference in average standard errors, empirical standard errors, and mean square errors between FCS and FCS with deletion, although both approaches were less precise than MVNI.

Table 5.1. Results for X_1 and X_2 binary, $\lambda = 2$, $RR(X_1, X_2) = 2$, and $\beta_1 = \beta_2 = \log(3)$.

Missing data mechanism	Outcome prevalence	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE	
Coordinated	0.10	MVNI	β_1	-0.08	0.30	0.28	0.951	0.08	
			β_2	-0.28	0.35	0.28	0.896	0.15	
		MVNI + deletion	β_1	0.06	0.30	0.31	0.955	0.10	
			β_2	-0.09	0.39	0.35	0.956	0.13	
		FCS	β_1	0.02	0.30	0.31	0.955	0.10	
			β_2	0.00	0.40	0.40	0.962	0.16	
	FCS + deletion	β_1	0.02	0.30	0.31	0.948	0.09		
		β_2	0.01	0.40	0.40	0.962	0.16		
	CCA	β_1	0.01	0.34	0.34	0.953	0.12		
		β_2	0.03	0.40	0.40	0.964	0.16		
	Coordinated	0.30	MVNI	β_1	0.03	0.16	0.15	0.952	0.02
				β_2	-0.32	0.17	0.16	0.547	0.13
MVNI + deletion			β_1	0.05	0.16	0.16	0.948	0.03	
			β_2	-0.15	0.20	0.19	0.872	0.06	
FCS			β_1	0.03	0.16	0.16	0.951	0.03	
			β_2	-0.11	0.20	0.21	0.893	0.05	
FCS + deletion		β_1	0.02	0.16	0.16	0.955	0.02		
		β_2	-0.06	0.21	0.21	0.932	0.05		
CCA		β_1	0.01	0.17	0.17	0.953	0.03		
		β_2	0.01	0.21	0.22	0.949	0.05		
Opposite		0.10	MVNI	β_1	-0.08	0.29	0.28	0.949	0.08
				β_2	-0.26	0.34	0.26	0.908	0.13
	MVNI + deletion		β_1	0.05	0.30	0.30	0.955	0.09	
			β_2	-0.07	0.37	0.33	0.964	0.11	
	FCS		β_1	0.01	0.30	0.31	0.952	0.10	
			β_2	0.03	0.39	0.39	0.963	0.16	
	FCS + deletion	β_1	0.01	0.30	0.31	0.950	0.09		
		β_2	0.05	0.39	0.40	0.964	0.16		
	CCA	β_1	0.03	0.39	0.41	0.956	0.17		
		β_2	0.03	0.39	0.39	0.965	0.15		
	Opposite	0.30	MVNI	β_1	0.00	0.16	0.15	0.961	0.02
				β_2	-0.20	0.18	0.16	0.805	0.06
MVNI + deletion			β_1	0.03	0.16	0.15	0.961	0.02	
			β_2	-0.02	0.20	0.19	0.952	0.03	
FCS			β_1	0.00	0.16	0.16	0.951	0.02	
			β_2	0.01	0.20	0.20	0.948	0.04	
FCS + deletion		β_1	-0.02	0.16	0.15	0.949	0.02		
		β_2	0.07	0.21	0.21	0.947	0.05		
CCA		β_1	0.01	0.20	0.20	0.952	0.04		
		β_2	0.02	0.20	0.20	0.952	0.04		

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Interestingly, CCA exhibited little bias in simulation scenarios involving categorical exposures, with a maximum absolute bias of 0.06 for both β_1 and β_2 . As expected, in discarding information from partially observed cases, CCA was noticeably less efficient than the MI approaches, especially for the coefficient β_1 for the fully observed exposure X_1 .

Simulation study 2: continuous exposures

To ensure that any deficiencies in performance in the continuous exposure setting could be attributed to the method of MI and not the use of modified Poisson regression for estimating relative risks, the accuracy of this method was first verified in complete datasets (i.e. before values in Y and X_2 were set to missing). Reassuringly, unbiased estimates for β_1 and β_2 were observed across all simulation scenarios (absolute bias ≤ 0.01), with estimated 95% confidence intervals demonstrating appropriate coverage (i.e. within the range 0.94 to 0.96) (results not shown).

The performance deficits of MI were more pronounced in the presence of continuous exposures than categorical exposures. Table 5.2 shows results for scenarios with a strong missing data mechanism ($\lambda = 2$), where $\text{corr}(X_1, X_2) = 0.70$ and $\beta_1 = \beta_2 = \log(3)$. A similar pattern of results was observed in other simulation scenarios, although absolute biases were smaller in magnitude for $\lambda = 1$ and $\beta_1 = \beta_2 = \log(2)$. As shown in Table 5.2, MVNI produced estimates for β_1 and β_2 that were biased towards the null, with the largest absolute bias observed for β_1 under the opposite mechanism with an outcome prevalence of 0.10 (relative risk estimate of 1.68 compared with the true value of 3). Across all 32 simulation scenarios, the median bias of MVNI was -0.21 for β_1 (range -0.58 to -0.10) and -0.12 for β_2 (range -0.27 to -0.06). Deleting imputed outcomes following MVNI reduced the bias of this imputation method, although moderate bias remained for β_2 in scenarios with an outcome prevalence of 0.30. The cost of this bias reduction was substantially larger average standard errors in comparison to MVNI. In terms of accuracy, deleting imputed outcomes following MVNI led to

reductions in the mean square error relative to MVNI without deletion in 26/32 and 12/32 simulation scenarios for β_1 and β_2 respectively.

FCS also produced estimates of β_1 and β_2 that were biased towards the null, albeit to a lesser degree than MVNI. The bias of -0.24 shown in Table 5.2 for an outcome prevalence of 0.30 under the coordinated mechanism translates to a relative risk estimate of just 2.37 versus the true value of 3. In addition to the more extreme simulation scenarios, noticeable bias for β_2 (absolute bias > 0.10) was apparent in simulation scenarios with an outcome prevalence of 0.10, a moderate missing data mechanism or where $\beta_1 = \beta_2 = \log(2)$. Deleting imputed outcomes following FCS tended to decrease the bias of this imputation approach, with absolute bias reduced in 28/32 and 26/32 simulation scenarios for β_1 and β_2 respectively. The trade-off for this bias reduction was a substantial loss in precision. Across the 32 simulation scenarios, average standard errors were 14.4% larger for β_1 and 8.1% larger for β_2 with the deletion of imputed outcomes following FCS compared to FCS alone. A consequence of the substantial loss in precision with the deletion of imputed outcomes following FCS was a loss in overall accuracy, with the mean square error increased relative to FCS without deletion in 30/32 and 26/32 simulation scenarios for β_1 and β_2 respectively.

Another noteworthy result from the continuous exposure setting was that average standard errors were consistently larger than empirical standard errors. Averaged across the 32 simulation scenarios, average standard errors for β_1 and β_2 were 25.8% and 17.9% larger than empirical standard errors respectively for MVNI, 14.4% and 11.9% larger for MVNI with deletion, 10.4% and 9.5% larger for FCS, and 14.3% and 12.1% larger for FCS with deletion. Discrepancies were most prominent in simulation scenarios with an outcome prevalence of 0.30. In scenarios where β_1 and β_2 were estimated with little bias, coverage probabilities also tended to be much higher than the nominal level of 0.95. Collectively these results suggest that estimated confidence intervals were too wide.

Table 5.2. Results for X_1 and X_2 continuous, $\lambda = 2$, $\text{Corr}(X_1, X_2) = 0.70$, and $\beta_1 = \beta_2 = \log(3)$.

Missing data mechanism	Outcome prevalence	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE	
Coordinated	0.10	MVNI	β_1	-0.56	0.48	0.39	0.838	0.47	
			β_2	-0.22	0.50	0.43	0.958	0.24	
		MVNI + deletion	β_1	0.01	0.58	0.55	0.965	0.30	
			β_2	-0.03	0.55	0.53	0.959	0.28	
		FCS	β_1	-0.08	0.51	0.49	0.961	0.25	
			β_2	-0.14	0.50	0.48	0.950	0.25	
	FCS + deletion	β_1	0.02	0.58	0.55	0.964	0.30		
		β_2	-0.04	0.55	0.53	0.961	0.28		
	CCA	β_1	0.01	0.66	0.67	0.943	0.45		
		β_2	0.01	0.53	0.55	0.936	0.30		
	Coordinated	0.30	MVNI	β_1	-0.26	0.26	0.20	0.890	0.11
				β_2	-0.27	0.25	0.20	0.859	0.11
MVNI + deletion			β_1	0.01	0.31	0.26	0.978	0.07	
			β_2	-0.11	0.28	0.23	0.963	0.07	
FCS			β_1	-0.09	0.26	0.22	0.962	0.06	
			β_2	-0.24	0.24	0.21	0.878	0.10	
FCS + deletion		β_1	0.02	0.31	0.26	0.980	0.07		
		β_2	-0.12	0.28	0.23	0.963	0.07		
CCA		β_1	0.02	0.32	0.32	0.951	0.11		
		β_2	0.00	0.25	0.26	0.950	0.07		
Opposite		0.10	MVNI	β_1	-0.58	0.47	0.37	0.830	0.47
				β_2	-0.17	0.48	0.42	0.961	0.21
	MVNI + deletion		β_1	0.00	0.56	0.52	0.966	0.28	
			β_2	0.02	0.53	0.51	0.959	0.26	
	FCS		β_1	-0.08	0.48	0.46	0.961	0.22	
			β_2	-0.07	0.49	0.47	0.959	0.22	
	FCS + deletion	β_1	0.01	0.56	0.52	0.971	0.27		
		β_2	0.01	0.53	0.51	0.962	0.26		
	CCA	β_1	0.00	0.60	0.62	0.939	0.39		
		β_2	0.01	0.48	0.50	0.938	0.25		
	Opposite	0.30	MVNI	β_1	-0.25	0.24	0.19	0.886	0.10
				β_2	-0.07	0.26	0.20	0.981	0.05
MVNI + deletion			β_1	0.01	0.30	0.25	0.983	0.06	
			β_2	0.06	0.29	0.23	0.980	0.06	
FCS			β_1	-0.07	0.24	0.21	0.974	0.05	
			β_2	-0.02	0.26	0.22	0.980	0.05	
FCS + deletion		β_1	0.02	0.29	0.25	0.983	0.06		
		β_2	0.05	0.29	0.23	0.982	0.06		
CCA		β_1	0.00	0.29	0.29	0.945	0.08		
		β_2	0.01	0.23	0.22	0.949	0.05		

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

As observed for categorical exposures, CCA exhibited little bias but tended to produce inefficient estimates of β_1 in scenarios involving continuous exposures. Interestingly, CCA produced more precise estimates of β_2 than the two MID approaches; across the 32 simulation scenarios, average standard errors for β_2 were 9.3% smaller with CCA relative to both deletion approaches.

Sensitivity analyses

In light of the relatively poor performance of the MI approaches for relative risk estimation, we undertook additional analyses to explore whether findings were sensitive to choices made during the fitting of imputation models or to the simulation parameters considered. First, we investigated the performance of simple rounding following MVNI as an alternative to adaptive rounding. While differences were minimal in most scenarios, MVNI introduced slightly more bias in both categorical and continuous exposure settings when simple rounding was used in place of adaptive rounding (results not shown). Next, we investigated the performance of FCS with the outcome imputed before rather than after the incomplete covariate X_2 . This modification made little difference to results (also not shown). We then explored the performance of the four MI approaches in scenarios involving $n = 250$ rather than $n = 1000$ observations. Excluding simulation scenarios with binary X_1 and X_2 where the reduced sample size resulted in zero cells in cross-tabulations involving the outcome (i.e. where log binomial analysis models would not converge), this change made little difference to the bias and coverage of parameter estimates (results not shown).

To investigate whether biased estimation would persist if the exposures were independent of one another, if the outcome was unrelated to one or both exposures, or if data were missing completely at random (i.e. probability of missing data unrelated to observed or unobserved data), several “null-case” simulation settings were considered. Table 5.3 shows results for continuous X_1 and X_2 under the coordinated missing data mechanism for an outcome prevalence of 0.30. The reference case for comparisons in this table was the previously

considered simulation scenario involving a strong missing data mechanism ($\lambda = 2$), where $\text{corr}(X_1, X_2) = 0.70$ and $\beta_1 = \beta_2 = \log(3)$. As shown in the table, the four MI approaches continued to produce biased parameter estimates when the exposures were independent of one another (i.e. $\text{corr}(X_1, X_2) = 0$). When the outcome was unrelated to one of the exposures, parameter estimates remained biased only for the exposure that was predictive of the outcome; little bias was observed with any of the MI approaches when both exposures were unrelated to the outcome. Lastly, bias was reduced but still evident when data were missing completely at random. A similar pattern of results was observed with binary X_1 and X_2 , and for an outcome prevalence of 0.10. Full results for these sensitivity analyses are available in web appendix C (see Section 5.3.6).

Table 5.3. Bias in scenarios with X_1 and X_2 continuous, coordinated missing data mechanism, and outcome prevalence = 0.30.

Simulation scenario	Parameter	MVNI	MVNI + deletion	FCS	FCS + deletion
1. $\text{Corr}(X_1, X_2) = 0.70$, $\beta_1 = \beta_2 = \log(3)$, $\lambda = 2$	β_1	-0.26	0.01	-0.09	0.02
	β_2	-0.27	-0.11	-0.24	-0.12
2. As in (1.), but with $\text{Corr}(X_1, X_2) = 0$	β_1	-0.27	-0.05	-0.15	-0.05
	β_2	-0.21	-0.06	-0.16	-0.06
3. As in (1.), but with $\beta_1 = 0$	β_1	-0.04	0.02	0.01	0.02
	β_2	-0.17	-0.03	-0.10	-0.03
4. As in (1.), but with $\beta_2 = 0$	β_1	-0.24	0.00	-0.10	0.00
	β_2	0.00	0.00	0.00	0.00
5. As in (1.), but with $\beta_1 = \beta_2 = 0$	β_1	-0.01	-0.01	-0.01	-0.01
	β_2	0.01	0.01	0.01	0.01
6. As in (1.), but with $\lambda = 0$ (MCAR)	β_1	-0.11	0.00	0.00	0.00
	β_2	-0.17	-0.08	-0.08	-0.08

Abbreviations: Corr, correlation; MCAR, missing completely at random; MVNI, multivariate normal imputation; FCS, fully conditional specification.

Lastly, to evaluate whether the performance deficiencies of FCS could be attributed solely to the misspecified logistic imputation model for the outcome, we considered additional simulation scenarios where missing data were restricted to either Y or X_2 only (with $\text{logit } P(\text{missing}) = \alpha + \lambda X_1$). Since data were missing in a single variable, missing values were imputed 20 times using logistic or linear regression as appropriate. Table 5.4 shows results for an outcome prevalence of

0.30, $\lambda = 2$ and $\beta_1 = \beta_2 = \log(3)$ for categorical exposures with $RR(X_1, X_2) = 2$ or continuous exposures with $\text{corr}(X_1, X_2) = 0.70$. The results for the original simulation scenario for FCS under the coordinated mechanism are also presented for comparison. As shown in the table, estimation remained biased when missing data were restricted to X_2 . Indeed for continuous exposures, the bias for β_2 was larger when missing data were restricted to X_2 compared to when missing data were restricted to Y . Thus it seems that the shortcomings of FCS were at least partly attributable to the choice of conditional imputation model for the incomplete covariate X_2 . It is worth noting that the bias following the imputation of continuous X_2 with a univariate linear model, as shown in Table 5.4, also suggests that the performance deficits seen with MVNI in the continuous exposure setting were partly due to inappropriate imputed values in the exposure (and not just the outcome).

Table 5.4. Bias in scenarios with $\lambda = 2$, outcome prevalence = 0.30, and $\beta_1 = \beta_2 = \log(3)$.

Simulation scenario	Parameter	Coordinated missing data in Y and X_2 (FCS)	Missing data in Y only	Missing data in X_2 only
Categorical X_1 and X_2 , $RR(X_1, X_2) = 2$	β_1	0.03	0.02	0.02
	β_2	-0.11	-0.06	-0.05
Continuous X_1 and X_2 , $\text{Corr}(X_1, X_2) = 0.70$	β_1	-0.09	-0.08	-0.03
	β_2	-0.24	-0.07	-0.22

Abbreviations: RR, relative risk; Corr, correlation; FCS, fully conditional specification.

5.3.5. Discussion

Given the widespread use of MI and the popularity of the relative risk, the lack of research on the application of MI for estimating the relative risk is surprising. In this study we demonstrated that standard model-based methods of MI can produce biased estimates of the relative risk with overly wide confidence intervals when data are MAR. Performance deficits were particularly evident when the analysis included continuous exposures, and in settings with larger relative risks, stronger missing data mechanisms and higher outcome prevalences. These findings raise concerns about the use of standard MI methods for relative risk estimation.

The primary aim of this study was to contrast the performance of MVNI and FCS for handling missing outcome data when estimating the relative risk. MVNI performed more poorly than FCS, producing relative risk estimates that were often substantially biased towards the null, both for categorical and continuous exposures. Although MVNI has been shown to be robust to violations of the multivariate normal assumption across a range of other settings, for example in estimating odds ratios or dealing with non-normal exposure variables (40, 41), such robustness to imputation model misspecification was not evident here. In contrast, FCS performed well when the analysis involved categorical exposures, only introducing noticeable bias for an outcome prevalence of 0.30, a strong missing data mechanism and large relative risks. Performance was less satisfactory in the presence of continuous exposures, with noticeable bias towards the null also evident in settings involving moderate relative risks or an outcome prevalence of 0.10. Even when relative risks for continuous exposures were estimated with little bias, FCS produced confidence intervals that were too wide. While we would recommend FCS over MVNI for relative risk estimation based on the simulation results presented here, clearly the approach is not without its shortcomings.

The secondary aim of this study was to evaluate whether deleting imputed outcomes improves the performance of MI for relative risk estimation. Focusing on FCS as the better performed method of MI, we observed little difference between FCS with and without deletion of imputed outcomes for analysis models involving categorical exposures. In the presence of continuous exposures, deleting imputed outcomes following FCS was associated with partial decreases in absolute bias at the expense of large increases in average standard errors; an interesting finding given that deletion improves the precision of estimation in settings where analysis and imputation models are the same (8). The lost precision with MID in the continuous exposure settings suggests that imputed values in the outcome contained information that was useful for analysis, which may be due to inconsistencies between the imputation and analysis models. Of course, since the imputation model was misspecified, this additional information could also result

in increased bias in a conventional MI analysis. In any case, we find it difficult to recommend MID for relative risk estimation based on these results, particularly since the approach is only advisable in settings where auxiliary variables for the outcome are unavailable (104).

Although logistic regression is the standard choice for imputing binary outcomes in software for implementing FCS, evidently this model is not optimal for relative risk estimation. Since controlling for confounding differs between the odds ratio and the relative risk (106), it is perhaps unsurprising that performance deficits were observed with FCS in this simulation study. This raises the question of whether an alternative conditional imputation model for the outcome should be adopted with FCS when relative risk estimation of interest. Assuming the analysis model is appropriately specified, an obvious candidate to minimise the problems of imputation model misspecification is the log binomial model, however issues with non-convergence could be a significant limitation in the context of FCS. As relative risks are often estimated using modified Poisson regression, another possibility would be to impute outcomes using Poisson regression. One difficulty with this approach is that imputed outcome values would be counts and would thus entail the use of modified Poisson regression in the analysis or the use of a rounding method prior to analysis with a log binomial model. Rounding methods have not been developed for this purpose. Another important challenge would be to incorporate a robust estimate of the error variance within the imputation model, since ordinary Poisson regression tends to overestimate the standard error for the relative risk (62). Although other approaches have been proposed to estimate relative risks (e.g. Cox regression with constant time at risk (63)), like Poisson regression, they typically require the use of a robust error variance which would need to be accounted for during imputation. This is difficult to achieve with current MI software.

In addition to the misspecified logistic model for imputing the outcome, sensitivity analyses revealed that the bias introduced by FCS could also be attributed to the conditional models used to impute the covariates. Imputing the

continuous covariate using linear regression in FCS assumed a linear relationship between the covariate and the outcome, which was inconsistent with the data generation model. A similar argument applies for the imputation of binary covariates using logistic regression. In a recent article, Bartlett and colleagues (30) proposed a modification to the standard FCS algorithm such that incomplete covariates are imputed from models that are compatible with the intended analysis model. While the approach seems promising in this context, further research is needed to understand its properties and suitability for relative risk estimation.

Due to convergence problems with the log binomial model in the continuous exposure setting, we elected to analyse all imputed datasets using the popular modified Poisson regression approach. Simulation results demonstrated that this method performed well in the absence of missing data, which is consistent with previous investigations of the method (62, 63, 105). An interesting consideration that arose following imputation was whether to use modified Poisson regression to analyse all imputed datasets or only those datasets where the log binomial model failed to converge. We chose the former approach, as this was simpler to implement and seemed more in keeping with Rubin's rules, however the latter could also be considered in future work.

Given the missing data mechanisms considered in the simulation study, it is not surprising that CCA produced parameter estimates with little bias. For missing data in a univariate outcome, CCA is known to produce unbiased and fully efficient of regression coefficients when the probability of missing data depends only on fully observed covariates (18, 23, 24). For missing data restricted to a covariate, CCA is known to be unbiased (but not fully efficient) if the probability of missing data is independent of the outcome conditional on the other covariates in the model (92). Both of these conditions were satisfied in the simulation study, where the probability of missing data in Y and X_2 depended only on the fully observed covariate X_1 . Clearly these conditions do not always hold in more complex practical settings, and CCA can introduce considerable bias when data are MAR. Taking into account the potential bias and inefficiency of CCA, we do

not advocate its use over MI for handling arbitrary patterns of missing data when estimating the relative risk.

Although we anticipate similar deficits with MVNI and FCS in more complex practical settings, it is difficult to draw definitive conclusions from a restricted set of simulation scenarios. Further exploration of the performance of these MI methods in real datasets (where the missing data mechanism is unknown) and in simulation scenarios with different covariate characteristics, outcome prevalences and missing data mechanisms would certainly be useful. A further limitation of the current study is that we did not evaluate alternatives to standard model-based methods of MI for handling missing data. Most notably we did not consider inverse probability weighting, a method that involves weighting complete cases in the analysis according to the inverse of the probability of being a complete case (26). We chose to focus on MI as it known to be more efficient than inverse probability weighting, particularly in the presence of auxiliary variables and for arbitrary patterns of missing data. However in light of the performance deficits of MI, further research could explore the use of inverse probability weighting in this setting. Within the MI framework, we did not consider less widely used model-based methods such as the general location model for mixtures of continuous and categorical variables, or non-parametric methods such as hot deck imputation. Again further research might consider the use of these approaches for relative risk estimation.

Conclusion

In summary, standard model-based methods of MI can produce biased and inefficient estimates of the relative risk due to misspecification of the imputation model. Should MI be chosen to handle missing data, we recommend researchers avoid MVNI and instead use FCS without deletion for estimating relative risks. However, further research is needed to identify optimal approaches for relative risk estimation within the MI framework.

5.3.6. Web appendix

Web appendix A. SAS code used in simulation studies.

Simulation study 1: categorical exposures

```
%macro categorical(seed, mvni_seed, fcs_seed, rrx, rry, intercept,
mechanism);

    *see note 1 for macro variable definitions;

    *1. Generate x1, x2 and y;

    data temp;
        length simulation id 5.;
        do simulation = 1 to 2000; *number of simulations = 2000;
            do id = 1 to 1000; *sample size = 1000;
                output;
            end;
        end;
    run;

    data temp;
        set temp;
        uniform1 = ranuni(&seed.);
        uniform2 = ranuni(&seed.);
        uniform3 = ranuni(&seed.);
        uniform4 = ranuni(&seed.);
        uniform5 = ranuni(&seed.);
    run;

    data temp;
        set temp;
        if uniform1 < 0.5 then x1 = 0;
        else x1 = 1;
    run;

    %if &rrx. = 2 %then %do;

        data temp;
            set temp;
            prob_x2 = (1/3) + (1/3)*x1;
        run;

    %end;

    %else %if &rrx. = 3 %then %do;

        data temp;
            set temp;
            prob_x2 = (1/4) + (2/4)*x1;
        run;

    %end;

    data temp;
        set temp;
        if uniform2 < prob_x2 then x2 = 1;
        else x2 = 0;
    run;

    data temp;
```

```

        set temp;
        prob_y = exp(&intercept. + log(&rry.)*x1 + log(&rry.)*x2);
        *solve intercept computationally to give desired outcome
        prevalence, see note 2 after macro;
run;

data temp;
    set temp;
    if uniform3 <= prob_y then y = 1;
    else y = 0;
run;

*2. Induce missing data in y and x2;

data temp;
    set temp;
    prob_y_strong = exp(-2.05 + 2*x1)/(1 + exp(-2.05 + 2*x1));
    *intercepts produce 30% missing data in y and 2;
    prob_y_mod = exp(-1.40 + 1*x1)/(1 + exp(-1.40 + 1*x1));
    prob_x2_strong = exp(-2.05 + 2*x1)/(1 + exp(-2.05 + 2*x1));
    prob_x2_mod = exp(-1.40 + 1*x1)/(1 + exp(-1.40 + 1*x1));
    prob_x2_opp_strong = exp(-0.05 - 2*x1)/(1 + exp(-0.05 - 2*x1));
    prob_x2_opp_mod = exp(-0.40 - 1*x1)/(1 + exp(-0.40 - 1*x1));
run;

%if &mechanism. = "Coordinated_strong" %then %do;

    data temp;
        set temp;
        if uniform4 <= prob_y_strong then missing_y = 1;
        else missing_y = 0;
        if uniform5 <= prob_x2_strong then missing_x2 = 1;
        else missing_x2 = 0;
    run;

%end;

%else %if &mechanism. = "Coordinated_mod" %then %do;

    data temp;
        set temp;
        if uniform4 <= prob_y_mod then missing_y = 1;
        else missing_y = 0;
        if uniform5 <= prob_x2_mod then missing_x2 = 1;
        else missing_x2 = 0;
    run;

%end;

%else %if &mechanism. = "Opposite_strong" %then %do;

    data temp;
        set temp;
        if uniform4 <= prob_y_strong then missing_y = 1;
        else missing_y = 0;
        if uniform5 <= prob_x2_opp_strong then missing_x2 = 1;
        else missing_x2 = 0;
    run;

%end;

%else %if &mechanism. = "Opposite_mod" %then %do;

    data temp;
        set temp;
        if uniform4 <= prob_y_mod then missing_y = 1;
        else missing_y = 0;
        if uniform5 <= prob_x2_opp_mod then missing_x2 = 1;
        else missing_x2 = 0;
    run;

%end;

```



```

run;

%end;

data temp;
  set temp;
  observed_y = y;
  if missing_y = 1 then observed_y = .;
  observed_x2 = x2;
  if missing_x2 = 1 then observed_x2 = .;
  keep simulation id x1 x2 y missing_y missing_x2 observed_y
  observed_x2;
run;

*3. Impute data;

*MVNI with adaptive rounding;

proc mi data=temp seed=&mvni_seed. nimpute=20 out=mvni;
  by simulation;
  mcmc chain=single initial=em;
  var observed_y observed_x2 x1;
run;

proc univariate data = mvni;
  by simulation _imputation_;
  var observed_y observed_x2;
  ods output BasicMeasures = bm;
run;

data temp1;
  set bm;
  where varname = 'observed_y' and locmeasure = 'Mean';
  rename locvalue = w_y;
  keep simulation _imputation_ locvalue;
run;

data temp2;
  set bm;
  where varname = 'observed_x2' and locmeasure = 'Mean';
  rename locvalue = w_x2;
  keep simulation _imputation_ locvalue;
run;

data mvni;
  merge mvni temp1 temp2;
  by simulation _imputation_;
  threshold_y = w_y - (quantile('NORMAL', w_y)*sqrt(w_y*(1-
  w_y)));
  threshold_x2 = w_x2 - (quantile('NORMAL', w_x2)*sqrt(w_x2*(1-
  w_x2)));
run;

data mvni;
  set mvni;
  if observed_y > threshold_y then observed_y = 1;
  else if observed_y <= threshold_y then observed_y = 0;
  if observed_x2 > threshold_x2 then observed_x2 = 1;
  else if observed_x2 <= threshold_x2 then observed_x2 = 0;
run;

*FCS;

proc mi data=temp seed=&fcs_seed. nimpute=20 out=fcs;
  by simulation;
  class observed_x2 observed_y x1;
  fcs logistic(observed_x2) logistic(observed_y) logistic(x1);
  var observed_x2 observed_y x1;
run;

```

```

%mend;

/*

Note 1: macro variable definitions

seed: starting seed for random number generation
mvni_seed: starting seed for multivariate normal imputation
fcs_seed: starting seed for fully conditional specification
rrx: relative risk for association between X1 and X2
rry: relative risk for association between Y and X1, and Y and X2
intercept: intercept value for generating P(Y=1) (see note 2 below)
mechanism: missing data mechanism, takes values "Coordinated_strong",
"Coordinated_mod", "Opposite_strong", "Opposite_mod" (see article for
mechanism definitions)

Note 2: values for macro variable <intercept> used in simulation study:

rry = 2, rrx = 2, prevalence = 0.10, intercept = -3.15
rry = 2, rrx = 3, prevalence = 0.10, intercept = -3.17
rry = 2, rrx = 2, prevalence = 0.30, intercept = -2.05
rry = 2, rrx = 3, prevalence = 0.30, intercept = -2.07
rry = 3, rrx = 2, prevalence = 0.10, intercept = -3.77
rry = 3, rrx = 3, prevalence = 0.10, intercept = -3.81
rry = 3, rrx = 2, prevalence = 0.30, intercept = -2.67
rry = 3, rrx = 3, prevalence = 0.30, intercept = -2.71

*/

*Example call of macro for rrx = 2, rry = 3, prevalence = 0.30, strong
coordinated mechanism;

%categorical(seed=1501, mvni_seed=1502, fcs_seed=1503, rrx=2, rry=3,
intercept=-2.67, mechanism = "Coordinated_strong");

*Example analysis with MVNI;

proc genmod data=mvni descending;
  class observed_y;
  by simulation_imputation_;
  model observed_y = observed_x2 x1 / link=log dist=binomial type3 wald
  covb;
  ods output parameterestimates = _estimates covb = _cov
  convergenceStatus = cs1; *check convergence status;
run;

data _cov;
  length parameter $11;
  set _cov;
  if rowname = 'Prm1' then Parameter = 'Intercept';
  if rowname = 'Prm2' then Parameter = 'observed_x2';
  if rowname = 'Prm3' then Parameter = 'x1';
  drop rowname;
  rename prm1 = Intercept prm2 = observed_x2 prm3 = x1;
run;

proc mianalyze parms=_estimates covb=_cov;
  by simulation;
  modeleffects Intercept observed_x2 x1;
  ods output parameterestimates=est1;
run;

```

Simulation study 2: continuous exposures

```
%macro continuous(seed, mvni_seed, fcs_seed, corrx, rry, intercept,
intercept_miss_y, intercept_miss_x2, mechanism);

    *see note 1 for macro variable definitions;

    *1. Generate x1, x2 and y;

    data temp;
        length simulation id 5.;
        do simulation = 1 to 2000; *number of simulations = 2000;
            do id = 1 to 1200;
                *sample size = 1000, allowing extra observations so that
                out of range values for P(Y=1) can be replaced;
                output;
            end;
        end;
    run;

    data temp;
        set temp;
        uniform1 = ranuni(&seed.);
        uniform2 = ranuni(&seed.);
        uniform3 = ranuni(&seed.);
        normal1 = rannor(&seed.);
        normal2 = rannor(&seed.);
    run;

    data temp;
        set temp;
        x1 = sqrt(0.2)*normal1;
    run;

    data temp;
        set temp;
        x2 = &corrx.*x1 + sqrt(0.2*(1-&corrx.*&corrx.))*normal2;
        x1_std = x1/sqrt(0.2);
    run;

    data temp;
        set temp;
        prob_y = exp(&intercept. + log(&rry.)*x1 + log(&rry.)*x2);
        *solve intercept computationally to give desired outcome
        prevalence, see note 2 after macro;
    run;

    data out_of_range;
        set temp;
        where prob_y > 1;
    run;

    data out_of_range;
        set out_of_range;
        count = _n_;
        indicator = 1;
    run;

    data out_of_range;
        set out_of_range;
        by indicator;
        if last.indicator;
        percent_out_range = 100*count/(1200*2000);
        *calculate percentage simulated values outside of range;
        keep percent_out_range;
    run;
```

```

data temp;
  set temp;
  if prob_y > 1 then delete; *exclude observations where
  probability out of range;
run;

data temp;
  retain counter;
  set temp;
  by simulation;
  if first.simulation then counter = 1;
  else counter = counter + 1;
run;

data temp;
  set temp;
  where counter <= 1000;
  if uniform1 <= prob_y then y = 1;
  else y = 0;
run;

*2. Induce missing data in y and x2;

%if &mechanism. = "Coordinated" %then %do;

  data temp;
    set temp;
    prob_y_miss = exp(&intercept_miss_y. + 2*x1_std)/(1 +
    exp(&intercept_miss_y. + 2*x1_std));
    prob_x2_miss = exp(&intercept_miss_x2. + 2*x1_std)/(1 +
    exp(&intercept_miss_x2. + 2*x1_std));
    *solve intercept computationally to produce 30% missing
    data in y and x2, see note 3 after macro;
  run;

%end;

%if &mechanism. = "Coordinated_mod" %then %do;

  data temp;
    set temp;
    prob_y_miss = exp(&intercept_miss_y. + 1*x1_std)/(1 +
    exp(&intercept_miss_y. + 1*x1_std));
    prob_x2_miss = exp(&intercept_miss_x2. + 1*x1_std)/(1 +
    exp(&intercept_miss_x2. + 1*x1_std));
  run;

%end;

%if &mechanism. = "Opposite" %then %do;

  data temp;
    set temp;
    prob_y_miss = exp(&intercept_miss_y. + 2*x1_std)/(1 +
    exp(&intercept_miss_y. + 2*x1_std));
    prob_x2_miss = exp(&intercept_miss_x2. - 2*x1_std)/(1 +
    exp(&intercept_miss_x2. - 2*x1_std));
  run;

%end;

%if &mechanism. = "Opposite_mod" %then %do;

  data temp;
    set temp;
    prob_y_miss = exp(&intercept_miss_y. + 1*x1_std)/(1 +
    exp(&intercept_miss_y. + 1*x1_std));
    prob_x2_miss = exp(&intercept_miss_x2. - 1*x1_std)/(1 +
    exp(&intercept_miss_x2. - 1*x1_std));

```

```

run;

%end;

data temp;
  set temp;
  if uniform2 <= prob_y_miss then missing_y = 1;
  else missing_y = 0;
  if uniform3 <= prob_x2_miss then missing_x2 = 1;
  else missing_x2 = 0;
run;

data temp;
  set temp;
  observed_y = y;
  if missing_y = 1 then observed_y = .;
  observed_x2 = x2;
  if missing_x2 = 1 then observed_x2 = .;
  keep simulation id x1 x2 y missing_y missing_x2 observed_y
  observed_x2;
run;

*3. Impute data;

*MVNI with adaptive rounding;

proc mi data=temp seed=&mvni_seed. nimpute=20 out=mvni;
  by simulation;
  mcmc chain=single initial=em;
  var observed_y observed_x2 x1;
run;

proc univariate data = mvni;
  by simulation _imputation_;
  var observed_y;
  ods output BasicMeasures = bm;
run;

data temp1;
  set bm;
  where varname = 'observed_y' and locmeasure = 'Mean';
  rename locvalue = w_y;
  keep simulation _imputation_ locvalue;
run;

data mvni;
  merge mvni temp1;
  by simulation _imputation_;
  threshold_y = w_y - (quantile('NORMAL', w_y)*sqrt(w_y*(1-
  w_y)));
run;

data mvni;
  set mvni;
  if observed_y > threshold_y then observed_y = 1;
  else if observed_y <= threshold_y then observed_y = 0;
run;

*FCS;

proc mi data=temp seed=&fcs_seed. nimpute=20 out=fcs;
  by simulation;
  class observed_y;
  fcs reg(observed_x2) logistic(observed_y) reg(x1);
  var observed_x2 observed_y x1;
run;

%mend;

```

```
/*
```

```
Note 1: macro variable definitions
```

```
seed: starting seed for random number generation  
mvni_seed: starting seed for multivariate normal imputation  
fcs_seed: starting seed for fully conditional specification  
corr_x: correlation for association between X1 and X2  
rry: relative risk for association between Y and X1, and Y and X2  
intercept: intercept value for generating P(Y=1) (see note 2 below)  
intercept_miss_y: intercept value for generating 30% missing data in Y (see  
note 3 below)  
intercept_miss_x2: intercept value for generating 30% missing data in X2 (see  
note 3 below)  
mechanism: missing data mechanism, takes values "Coordinated_strong",  
"Coordinated_mod", "Opposite_strong", "Opposite_mod" (see article for  
mechanism definitions)
```

```
Note 2: values for macro variable <intercept> used in simulation study:
```

```
rry = 2, corr_x = 0.30, prevalence = 0.10, intercept = -2.43  
rry = 2, corr_x = 0.70, prevalence = 0.10, intercept = -2.46  
rry = 2, corr_x = 0.30, prevalence = 0.30, intercept = -1.32  
rry = 2, corr_x = 0.70, prevalence = 0.30, intercept = -1.34  
rry = 3, corr_x = 0.30, prevalence = 0.10, intercept = -2.61  
rry = 3, corr_x = 0.70, prevalence = 0.10, intercept = -2.69  
rry = 3, corr_x = 0.30, prevalence = 0.30, intercept = -1.37  
rry = 3, corr_x = 0.70, prevalence = 0.30, intercept = -1.36
```

```
Note 3: values for macro variables <intercept_miss_y> and <intercept_miss_x2>  
used in simulation study:
```

```
Moderate mechanism, intercept_miss_y = intercept_miss_x2 = -1.02  
Strong mechanism, intercept_miss_y = intercept_miss_x2 = -1.39
```

```
-> Note intercept values adjusted in scenarios with an outcome prevalence of  
0.30 to maintain 30% missing data after replacing out of range values for  
P(Y=1)
```

```
*/
```

```
*Example call of macro for corr_x = 0.70, rry = 2, prevalence = 0.10, moderate  
coordinated mechanism;
```

```
%continuous (seed=12601, mvni_seed=12602, fcs_seed=12603, corr_x=0.70, rry=2,  
intercept=-2.46, intercept_miss_y = -1.02, intercept_miss_x2 = -1.02,  
mechanism = "Coordinated_mod");
```

```
*Example analysis with FCS (using log Poisson GEE to estimate the relative  
risk);
```

```
proc genmod data=fcs;  
  class id;  
  by simulation _imputation_;  
  model observed_y = observed_x2 x1 / link=log dist=poisson type3 wald  
  covb;  
  repeated subject = id /type=ind;  
  ods output GEEempest = _estimates covb = _cov;  
run;
```

```
data _estimates;  
  set _estimates;  
  rename parm = parameter;  
  keep simulation _imputation_ Parm Estimate stderr;  
run;
```

```
data _cov;
    length parameter $11;
    set _cov;
    if rowname = 'Prm1' then Parameter = 'Intercept';
    if rowname = 'Prm2' then Parameter = 'observed_x2';
    if rowname = 'Prm3' then Parameter = 'x1';
    drop rowname;
    rename prm1 = Intercept prm2 = observed_x2 prm3 = x1;
run;

proc mianalyze parms=_estimates covb=_cov;
    by simulation;
    modeleffects Intercept observed_x2 x1;
    ods output parameterestimates=est3;
run;
```

Web appendix B. Full results from simulation studies 1 and 2.

Table 5.5. Simulation results for X_1 and X_2 binary, coordinated mechanism, $\lambda = 1$.

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.10	Log(2)	2	MVNI	β_1	-0.04	0.26	0.24	0.961	0.06			
				β_2	-0.14	0.30	0.26	0.941	0.09			
			MVNI + deletion	β_1	0.05	0.27	0.27	0.948	0.08			
				β_2	-0.05	0.32	0.30	0.960	0.10			
			FCS	β_1	0.02	0.27	0.28	0.945	0.08			
				β_2	0.00	0.33	0.34	0.957	0.11			
			FCS + deletion	β_1	0.02	0.27	0.28	0.947	0.08			
				β_2	0.01	0.33	0.34	0.957	0.11			
			CCA	β_1	0.02	0.31	0.32	0.954	0.10			
				β_2	0.01	0.33	0.33	0.957	0.11			
			0.10	Log(2)	3	MVNI	β_1	-0.04	0.29	0.27	0.968	0.07
							β_2	-0.14	0.33	0.28	0.957	0.10
MVNI + deletion	β_1	0.05				0.30	0.30	0.958	0.09			
	β_2	-0.05				0.36	0.33	0.962	0.11			
FCS	β_1	0.02				0.31	0.31	0.955	0.10			
	β_2	0.01				0.37	0.37	0.947	0.14			
FCS + deletion	β_1	0.01				0.30	0.31	0.952	0.10			
	β_2	0.02				0.37	0.37	0.951	0.14			
CCA	β_1	0.00				0.34	0.34	0.953	0.12			
	β_2	0.03				0.36	0.37	0.948	0.14			
0.10	Log(3)	2				MVNI	β_1	-0.07	0.30	0.27	0.958	0.08
							β_2	-0.27	0.34	0.27	0.895	0.15
			MVNI + deletion	β_1	0.08	0.30	0.32	0.951	0.11			
				β_2	-0.10	0.38	0.35	0.944	0.13			
			FCS	β_1	0.04	0.31	0.32	0.953	0.10			
				β_2	0.02	0.40	0.41	0.954	0.17			
			FCS + deletion	β_1	0.03	0.30	0.32	0.953	0.10			
				β_2	0.03	0.40	0.41	0.958	0.17			
			CCA	β_1	0.04	0.35	0.36	0.953	0.13			
				β_2	0.03	0.39	0.41	0.951	0.17			
			0.10	Log(3)	3	MVNI	β_1	-0.04	0.34	0.30	0.969	0.09
							β_2	-0.29	0.37	0.30	0.901	0.17
MVNI + deletion	β_1	0.11				0.35	0.35	0.957	0.13			
	β_2	-0.11				0.42	0.38	0.954	0.16			
FCS	β_1	0.05				0.35	0.35	0.955	0.13			
	β_2	0.02				0.44	0.45	0.957	0.21			
FCS + deletion	β_1	0.05				0.35	0.35	0.948	0.13			
	β_2	0.03				0.44	0.45	0.958	0.21			
CCA	β_1	0.04				0.39	0.39	0.954	0.16			
	β_2	0.04				0.44	0.45	0.959	0.21			

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.30	Log(2)	2	MVNI	β_1	-0.01	0.14	0.13	0.959	0.02			
				β_2	-0.13	0.15	0.14	0.884	0.04			
			MVNI + deletion	β_1	0.03	0.14	0.14	0.950	0.02			
				β_2	-0.07	0.17	0.15	0.941	0.03			
			FCS	β_1	0.01	0.14	0.14	0.949	0.02			
				β_2	-0.03	0.17	0.17	0.954	0.03			
			FCS + deletion	β_1	0.01	0.14	0.14	0.950	0.02			
				β_2	-0.01	0.17	0.17	0.953	0.03			
			CCA	β_1	0.01	0.16	0.16	0.949	0.02			
				β_2	0.00	0.17	0.17	0.957	0.03			
			0.30	Log(2)	3	MVNI	β_1	0.01	0.15	0.15	0.958	0.02
							β_2	-0.14	0.17	0.16	0.881	0.04
MVNI + deletion	β_1	0.05				0.16	0.15	0.949	0.03			
	β_2	-0.07				0.18	0.17	0.939	0.03			
FCS	β_1	0.02				0.16	0.16	0.953	0.02			
	β_2	-0.02				0.19	0.19	0.950	0.04			
FCS + deletion	β_1	0.01				0.16	0.16	0.949	0.02			
	β_2	-0.01				0.19	0.19	0.952	0.04			
CCA	β_1	0.00				0.17	0.17	0.948	0.03			
	β_2	0.01				0.19	0.19	0.951	0.04			
0.30	Log(3)	2				MVNI	β_1	-0.01	0.16	0.15	0.955	0.02
							β_2	-0.26	0.17	0.15	0.682	0.09
			MVNI + deletion	β_1	0.05	0.16	0.16	0.948	0.03			
				β_2	-0.12	0.19	0.18	0.893	0.05			
			FCS	β_1	0.01	0.16	0.16	0.954	0.03			
				β_2	-0.06	0.20	0.20	0.933	0.04			
			FCS + deletion	β_1	0.01	0.16	0.16	0.951	0.03			
				β_2	-0.03	0.20	0.20	0.946	0.04			
			CCA	β_1	0.01	0.18	0.18	0.954	0.03			
				β_2	0.01	0.20	0.20	0.955	0.04			
			0.30	Log(3)	3	MVNI	β_1	0.03	0.18	0.18	0.948	0.03
							β_2	-0.27	0.19	0.18	0.719	0.10
MVNI + deletion	β_1	0.08				0.18	0.19	0.931	0.04			
	β_2	-0.13				0.22	0.20	0.901	0.06			
FCS	β_1	0.03				0.19	0.19	0.940	0.04			
	β_2	-0.06				0.23	0.22	0.934	0.05			
FCS + deletion	β_1	0.02				0.18	0.19	0.938	0.04			
	β_2	-0.02				0.23	0.23	0.945	0.05			
CCA	β_1	0.01				0.20	0.21	0.942	0.04			
	β_2	0.01				0.23	0.23	0.950	0.05			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.6. Simulation results for X_1 and X_2 binary, coordinated mechanism, $\lambda = 2$.

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE	
0.10	Log(2)	2	MVNI	β_1	-0.07	0.26	0.24	0.954	0.06	
				β_2	-0.15	0.30	0.26	0.936	0.09	
			MVNI + deletion	β_1	0.03	0.27	0.27	0.956	0.07	
				β_2	-0.05	0.33	0.31	0.958	0.10	
			FCS	β_1	0.01	0.27	0.27	0.947	0.07	
				β_2	-0.00	0.34	0.34	0.957	0.12	
	FCS + deletion	β_1	0.01	0.27	0.27	0.950	0.07			
		β_2	0.00	0.34	0.34	0.955	0.12			
	CCA	β_1	-0.01	0.32	0.32	0.954	0.10			
		β_2	0.01	0.34	0.34	0.953	0.11			
	0.10	Log(2)	3	MVNI	β_1	-0.05	0.29	0.28	0.951	0.08
					β_2	-0.14	0.33	0.29	0.948	0.11
MVNI + deletion				β_1	0.05	0.30	0.31	0.947	0.10	
				β_2	-0.04	0.36	0.34	0.956	0.12	
FCS				β_1	0.02	0.30	0.31	0.942	0.10	
				β_2	0.00	0.37	0.37	0.949	0.14	
FCS + deletion		β_1	0.01	0.30	0.31	0.943	0.10			
		β_2	0.01	0.37	0.37	0.950	0.14			
CCA		β_1	-0.01	0.35	0.36	0.952	0.13			
		β_2	0.02	0.36	0.37	0.948	0.14			
0.10		Log(3)	2	MVNI	β_1	-0.08	0.30	0.28	0.951	0.08
					β_2	-0.28	0.35	0.28	0.896	0.15
	MVNI + deletion			β_1	0.06	0.30	0.31	0.955	0.10	
				β_2	-0.09	0.39	0.35	0.956	0.13	
	FCS			β_1	0.02	0.30	0.31	0.955	0.10	
				β_2	-0.00	0.40	0.40	0.962	0.16	
	FCS + deletion	β_1	0.02	0.30	0.31	0.948	0.09			
		β_2	0.01	0.40	0.40	0.962	0.16			
	CCA	β_1	0.01	0.34	0.34	0.953	0.12			
		β_2	0.03	0.40	0.40	0.964	0.16			
	0.10	Log(3)	3	MVNI	β_1	-0.03	0.34	0.32	0.961	0.10
					β_2	-0.28	0.39	0.32	0.893	0.18
MVNI + deletion				β_1	0.10	0.34	0.35	0.955	0.13	
				β_2	-0.10	0.43	0.39	0.955	0.16	
FCS				β_1	0.05	0.35	0.35	0.947	0.13	
				β_2	-0.00	0.44	0.45	0.956	0.20	
FCS + deletion		β_1	0.04	0.34	0.35	0.949	0.12			
		β_2	0.02	0.44	0.45	0.960	0.20			
CCA		β_1	0.03	0.38	0.40	0.951	0.16			
		β_2	0.04	0.44	0.45	0.963	0.20			
0.30		Log(2)	2	MVNI	β_1	0.00	0.14	0.14	0.949	0.02
					β_2	-0.16	0.16	0.14	0.836	0.04
	MVNI + deletion			β_1	0.03	0.14	0.14	0.939	0.02	

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.30	Log(2)	3	FCS	β_2	-0.07	0.17	0.16	0.933	0.03			
				β_1	0.01	0.14	0.14	0.943	0.02			
			FCS + deletion	β_2	-0.05	0.17	0.17	0.940	0.03			
				β_1	0.01	0.14	0.14	0.946	0.02			
			CCA	β_2	-0.02	0.17	0.17	0.949	0.03			
				β_1	-0.00	0.16	0.16	0.945	0.03			
				β_2	0.00	0.17	0.17	0.953	0.03			
			0.30	Log(2)	3	MVNI	β_1	0.03	0.15	0.15	0.955	0.02
							β_2	-0.16	0.17	0.16	0.843	0.05
						MVNI + deletion	β_1	0.05	0.15	0.15	0.949	0.03
							β_2	-0.07	0.18	0.18	0.927	0.04
						FCS	β_1	0.02	0.16	0.16	0.948	0.03
β_2	-0.05	0.19					0.19	0.930	0.04			
FCS + deletion	β_1	0.01				0.15	0.16	0.945	0.02			
	β_2	-0.02				0.19	0.19	0.940	0.04			
CCA	β_1	0.00				0.17	0.17	0.954	0.03			
	β_2	0.00				0.19	0.20	0.941	0.04			
0.30	Log(3)	2				MVNI	β_1	0.03	0.16	0.15	0.952	0.02
							β_2	-0.32	0.17	0.16	0.547	0.13
			MVNI + deletion	β_1	0.05	0.16	0.16	0.948	0.03			
				β_2	-0.15	0.20	0.19	0.872	0.06			
			FCS	β_1	0.03	0.16	0.16	0.951	0.03			
				β_2	-0.11	0.20	0.21	0.893	0.05			
			FCS + deletion	β_1	0.02	0.16	0.16	0.955	0.02			
				β_2	-0.06	0.21	0.21	0.932	0.05			
			CCA	β_1	0.01	0.17	0.17	0.953	0.03			
				β_2	0.01	0.21	0.22	0.949	0.05			
			0.30	Log(3)	3	MVNI	β_1	0.07	0.18	0.18	0.938	0.04
							β_2	-0.31	0.20	0.18	0.640	0.13
MVNI + deletion	β_1	0.07				0.18	0.18	0.939	0.04			
	β_2	-0.14				0.22	0.21	0.895	0.06			
FCS	β_1	0.04				0.18	0.18	0.946	0.04			
	β_2	-0.10				0.23	0.22	0.921	0.06			
FCS + deletion	β_1	0.02				0.18	0.18	0.949	0.03			
	β_2	-0.04				0.23	0.23	0.939	0.06			
CCA	β_1	-0.00				0.19	0.20	0.952	0.04			
	β_2	0.02				0.23	0.23	0.942	0.06			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.7. Simulation results for X_1 and X_2 binary, opposite mechanism, $\lambda = 1$.

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE	
0.10	Log(2)	2	MVNI	β_1	-0.05	0.26	0.23	0.964	0.06	
				β_2	-0.13	0.30	0.25	0.953	0.08	
			MVNI + deletion	β_1	0.04	0.27	0.26	0.958	0.07	
				β_2	-0.03	0.32	0.30	0.961	0.09	
			FCS	β_1	0.01	0.27	0.27	0.953	0.07	
				β_2	0.03	0.33	0.34	0.950	0.12	
	FCS + deletion	β_1	0.01	0.27	0.27	0.954	0.07			
		β_2	0.03	0.33	0.34	0.950	0.12			
	0.10	Log(2)	3	MVNI	β_1	-0.04	0.29	0.26	0.969	0.07
					β_2	-0.14	0.32	0.28	0.950	0.10
				MVNI + deletion	β_1	0.05	0.30	0.29	0.961	0.09
					β_2	-0.04	0.35	0.33	0.957	0.11
FCS				β_1	0.00	0.31	0.30	0.951	0.09	
				β_2	0.04	0.37	0.38	0.949	0.15	
FCS + deletion	β_1	-0.00	0.30	0.30	0.951	0.09				
	β_2	0.04	0.36	0.38	0.946	0.15				
0.10	Log(3)	2	MVNI	β_1	-0.08	0.30	0.26	0.956	0.08	
				β_2	-0.24	0.33	0.26	0.907	0.13	
			MVNI + deletion	β_1	0.07	0.30	0.31	0.952	0.10	
				β_2	-0.06	0.37	0.34	0.959	0.12	
			FCS	β_1	0.02	0.31	0.32	0.948	0.10	
				β_2	0.06	0.39	0.41	0.953	0.17	
FCS + deletion	β_1	0.02	0.30	0.31	0.946	0.10				
	β_2	0.07	0.39	0.41	0.953	0.17				
0.10	Log(3)	3	MVNI	β_1	-0.04	0.33	0.29	0.973	0.09	
				β_2	-0.27	0.36	0.28	0.907	0.15	
			MVNI + deletion	β_1	0.10	0.35	0.35	0.955	0.13	
				β_2	-0.09	0.40	0.36	0.957	0.14	
			FCS	β_1	0.03	0.35	0.36	0.948	0.13	
				β_2	0.05	0.43	0.45	0.955	0.21	
FCS + deletion	β_1	0.03	0.35	0.36	0.945	0.13				
	β_2	0.06	0.43	0.45	0.958	0.21				
0.10	Log(3)	3	CCA	β_1	0.06	0.38	0.69	0.950	0.48	
				β_2	0.06	0.38	0.41	0.952	0.17	
			MVNI	β_1	-0.02	0.14	0.13	0.951	0.02	
				β_2	-0.10	0.15	0.14	0.905	0.03	
			MVNI + deletion	β_1	0.02	0.14	0.14	0.945	0.02	
				β_2	0.05	0.43	0.45	0.959	0.20	

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
			FCS	β_2	-0.03	0.16	0.16	0.943	0.03			
				β_1	-0.00	0.14	0.14	0.946	0.02			
				β_2	0.00	0.17	0.17	0.947	0.03			
				FCS + deletion	β_1	-0.00	0.14	0.14	0.946	0.02		
					β_2	0.02	0.17	0.17	0.944	0.03		
				CCA	β_1	0.01	0.17	0.17	0.949	0.03		
			β_2		0.01	0.17	0.17	0.944	0.03			
			0.30	Log(2)	3	MVNI	β_1	0.00	0.15	0.14	0.964	0.02
							β_2	-0.11	0.17	0.15	0.901	0.04
						MVNI + deletion	β_1	0.04	0.16	0.15	0.949	0.02
							β_2	-0.04	0.18	0.17	0.945	0.03
						FCS	β_1	0.00	0.16	0.16	0.956	0.02
β_2	0.01	0.19					0.19	0.947	0.03			
FCS + deletion	β_1	-0.00				0.16	0.16	0.952	0.02			
	β_2	0.02				0.19	0.19	0.951	0.04			
CCA	β_1	0.01				0.18	0.18	0.949	0.03			
	β_2	0.01				0.18	0.18	0.948	0.03			
0.30	Log(3)	2				MVNI	β_1	-0.03	0.16	0.15	0.955	0.02
							β_2	-0.19	0.17	0.16	0.798	0.06
			MVNI + deletion	β_1	0.03	0.16	0.15	0.958	0.02			
				β_2	-0.05	0.19	0.18	0.942	0.04			
			FCS	β_1	-0.01	0.16	0.16	0.952	0.02			
				β_2	0.00	0.20	0.20	0.954	0.04			
			FCS + deletion	β_1	-0.01	0.16	0.16	0.947	0.02			
				β_2	0.04	0.20	0.20	0.952	0.04			
			CCA	β_1	0.00	0.19	0.20	0.949	0.04			
				β_2	0.01	0.19	0.20	0.955	0.04			
			0.30	Log(3)	3	MVNI	β_1	0.01	0.18	0.17	0.960	0.03
							β_2	-0.21	0.19	0.18	0.810	0.07
MVNI + deletion	β_1	0.05				0.18	0.18	0.951	0.04			
	β_2	-0.07				0.21	0.20	0.944	0.05			
FCS	β_1	-0.00				0.19	0.19	0.947	0.04			
	β_2	0.01				0.23	0.23	0.949	0.05			
FCS + deletion	β_1	-0.01				0.18	0.19	0.944	0.04			
	β_2	0.05				0.23	0.23	0.949	0.06			
CCA	β_1	0.01				0.22	0.23	0.943	0.05			
	β_2	0.02				0.22	0.23	0.946	0.05			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.8. Simulation results for X_1 and X_2 binary, opposite mechanism, $\lambda = 2$.

Outcome prevalence	β_1, β_2	RR(X_1, X_2)	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE	
0.10	Log(2)	2	MVNI	β_1	-0.06	0.26	0.24	0.963	0.06	
				β_2	-0.13	0.31	0.27	0.951	0.09	
			MVNI + deletion	β_1	0.04	0.27	0.27	0.959	0.07	
				β_2	-0.03	0.33	0.32	0.960	0.10	
			FCS	β_1	0.01	0.27	0.27	0.959	0.07	
				β_2	0.03	0.34	0.36	0.955	0.13	
	FCS + deletion	β_1	0.01	0.27	0.27	0.959	0.07			
		β_2	0.03	0.34	0.36	0.946	0.13			
	CCA	β_1	0.02	0.34	0.34	0.962	0.12			
		β_2	0.02	0.34	0.36	0.954	0.13			
	0.10	Log(2)	3	MVNI	β_1	-0.05	0.29	0.27	0.959	0.08
					β_2	-0.14	0.33	0.28	0.952	0.10
MVNI + deletion				β_1	0.05	0.30	0.30	0.949	0.09	
				β_2	-0.04	0.36	0.33	0.960	0.11	
FCS				β_1	0.01	0.31	0.31	0.943	0.10	
				β_2	0.04	0.38	0.39	0.948	0.15	
FCS + deletion		β_1	-0.00	0.30	0.31	0.940	0.10			
		β_2	0.05	0.38	0.39	0.949	0.16			
CCA		β_1	0.03	0.37	0.39	0.954	0.15			
		β_2	0.04	0.37	0.39	0.947	0.15			
0.10		Log(3)	2	MVNI	β_1	-0.08	0.29	0.28	0.949	0.08
					β_2	-0.26	0.34	0.26	0.908	0.13
	MVNI + deletion			β_1	0.05	0.30	0.30	0.955	0.09	
				β_2	-0.07	0.37	0.33	0.964	0.11	
	FCS			β_1	0.01	0.30	0.31	0.952	0.10	
				β_2	0.03	0.39	0.39	0.963	0.16	
	FCS + deletion	β_1	0.01	0.30	0.31	0.950	0.09			
		β_2	0.05	0.39	0.40	0.964	0.16			
	CCA	β_1	0.03	0.39	0.41	0.956	0.17			
		β_2	0.03	0.39	0.39	0.965	0.15			
	0.10	Log(3)	3	MVNI	β_1	-0.04	0.33	0.30	0.967	0.09
					β_2	-0.27	0.38	0.29	0.906	0.16
MVNI + deletion				β_1	0.09	0.34	0.32	0.965	0.11	
				β_2	-0.09	0.41	0.37	0.952	0.14	
FCS				β_1	0.02	0.34	0.34	0.956	0.12	
				β_2	0.05	0.44	0.46	0.956	0.22	
FCS + deletion		β_1	0.01	0.34	0.34	0.956	0.11			
		β_2	0.07	0.44	0.47	0.957	0.22			
CCA		β_1	0.05	0.44	0.45	0.961	0.20			
		β_2	0.05	0.44	0.46	0.954	0.21			
0.30		Log(2)	2	MVNI	β_1	-0.01	0.14	0.13	0.955	0.02
					β_2	-0.10	0.16	0.14	0.916	0.03
	MVNI + deletion			β_1	0.02	0.14	0.14	0.952	0.02	

Outcome prevalence	β_1, β_2	$RR(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
			FCS	β_2	-0.02	0.17	0.16	0.959	0.03			
				β_1	-0.01	0.14	0.14	0.949	0.02			
			FCS + deletion	β_2	0.01	0.17	0.17	0.953	0.03			
				β_1	-0.01	0.14	0.14	0.946	0.02			
			CCA	β_2	0.03	0.18	0.18	0.955	0.03			
				β_1	0.00	0.17	0.18	0.951	0.03			
0.30	Log(2)	3	MVNI	β_1	0.01	0.15	0.15	0.952	0.02			
				β_2	-0.11	0.17	0.16	0.914	0.04			
			MVNI + deletion	β_1	0.03	0.15	0.15	0.950	0.02			
				β_2	-0.03	0.19	0.18	0.953	0.03			
			FCS	β_1	-0.01	0.16	0.16	0.941	0.03			
				β_2	0.01	0.19	0.20	0.954	0.04			
			FCS + deletion	β_1	-0.02	0.16	0.16	0.938	0.03			
				β_2	0.04	0.19	0.20	0.948	0.04			
			CCA	β_1	0.00	0.19	0.20	0.935	0.04			
				β_2	0.01	0.19	0.19	0.951	0.04			
			0.30	Log(3)	2	MVNI	β_1	-0.00	0.16	0.15	0.961	0.02
							β_2	-0.20	0.18	0.16	0.805	0.06
MVNI + deletion	β_1	0.03				0.16	0.15	0.961	0.02			
	β_2	-0.02				0.20	0.19	0.952	0.03			
FCS	β_1	-0.00				0.16	0.16	0.951	0.02			
	β_2	0.01				0.20	0.20	0.948	0.04			
FCS + deletion	β_1	-0.02				0.16	0.15	0.949	0.02			
	β_2	0.07				0.21	0.21	0.947	0.05			
CCA	β_1	0.01				0.20	0.20	0.952	0.04			
	β_2	0.02				0.20	0.20	0.952	0.04			
0.30	Log(3)	3				MVNI	β_1	0.04	0.18	0.17	0.961	0.03
							β_2	-0.22	0.20	0.18	0.809	0.08
			MVNI + deletion	β_1	0.05	0.18	0.17	0.961	0.03			
				β_2	-0.04	0.22	0.21	0.946	0.05			
			FCS	β_1	-0.00	0.18	0.18	0.956	0.03			
				β_2	0.01	0.23	0.24	0.948	0.06			
			FCS + deletion	β_1	-0.02	0.18	0.18	0.951	0.03			
				β_2	0.07	0.24	0.24	0.942	0.06			
			CCA	β_1	0.02	0.23	0.22	0.956	0.05			
				β_2	0.01	0.23	0.23	0.949	0.05			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.9. Simulation results for X_1 and X_2 continuous, coordinated mechanism, $\lambda = 1$.

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.10	Log(2)	0.30	MVNI	β_1	-0.18	0.29	0.24	0.952	0.09			
				β_2	-0.11	0.33	0.29	0.963	0.10			
			MVNI + deletion	β_1	-0.01	0.33	0.31	0.962	0.10			
				β_2	-0.01	0.35	0.34	0.958	0.12			
			FCS	β_1	-0.02	0.31	0.30	0.958	0.09			
				β_2	-0.02	0.34	0.33	0.955	0.11			
			FCS + deletion	β_1	-0.00	0.33	0.31	0.964	0.10			
				β_2	-0.01	0.35	0.34	0.961	0.11			
			CCA	β_1	-0.00	0.39	0.39	0.935	0.15			
				β_2	0.01	0.34	0.34	0.940	0.12			
			0.10	Log(2)	0.70	MVNI	β_1	-0.20	0.42	0.36	0.953	0.17
							β_2	-0.11	0.45	0.40	0.971	0.17
MVNI + deletion	β_1	0.01				0.47	0.45	0.963	0.20			
	β_2	-0.00				0.49	0.47	0.958	0.22			
FCS	β_1	-0.01				0.44	0.43	0.953	0.19			
	β_2	-0.03				0.47	0.46	0.955	0.21			
FCS + deletion	β_1	0.01				0.47	0.45	0.959	0.20			
	β_2	-0.01				0.49	0.47	0.963	0.22			
CCA	β_1	0.00				0.51	0.52	0.935	0.27			
	β_2	0.01				0.47	0.48	0.943	0.23			
0.10	Log(3)	0.30				MVNI	β_1	-0.29	0.30	0.24	0.892	0.15
							β_2	-0.21	0.34	0.29	0.943	0.13
			MVNI + deletion	β_1	-0.01	0.35	0.32	0.969	0.10			
				β_2	-0.04	0.37	0.35	0.964	0.12			
			FCS	β_1	-0.05	0.32	0.30	0.963	0.09			
				β_2	-0.09	0.35	0.33	0.953	0.12			
			FCS + deletion	β_1	-0.01	0.35	0.32	0.972	0.10			
				β_2	-0.04	0.37	0.35	0.958	0.12			
			CCA	β_1	0.01	0.41	0.42	0.938	0.18			
				β_2	-0.01	0.35	0.36	0.940	0.13			
			0.10	Log(3)	0.70	MVNI	β_1	-0.31	0.45	0.37	0.939	0.23
							β_2	-0.23	0.47	0.40	0.963	0.22
MVNI + deletion	β_1	0.03				0.50	0.47	0.970	0.22			
	β_2	-0.06				0.52	0.48	0.970	0.23			
FCS	β_1	-0.02				0.46	0.44	0.965	0.19			
	β_2	-0.12				0.48	0.45	0.965	0.22			
FCS + deletion	β_1	0.03				0.50	0.47	0.971	0.22			
	β_2	-0.06				0.52	0.48	0.969	0.23			
CCA	β_1	0.02				0.55	0.56	0.940	0.32			
	β_2	0.00				0.50	0.51	0.942	0.26			
0.30	Log(2)	0.30				MVNI	β_1	-0.11	0.16	0.13	0.956	0.03
							β_2	-0.10	0.18	0.14	0.956	0.03
			MVNI + deletion	β_1	-0.00	0.19	0.15	0.987	0.02			

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.30	Log(2)	0.70	FCS	β_2	-0.02	0.19	0.16	0.981	0.03			
				β_1	-0.04	0.17	0.14	0.975	0.02			
			FCS + deletion	β_2	-0.06	0.18	0.15	0.965	0.03			
				β_1	-0.00	0.19	0.15	0.985	0.02			
			CCA	β_2	-0.02	0.19	0.16	0.983	0.03			
				β_1	0.00	0.20	0.19	0.953	0.04			
				β_2	0.01	0.17	0.17	0.950	0.03			
			0.30	Log(2)	0.70	MVNI	β_1	-0.10	0.23	0.19	0.968	0.05
							β_2	-0.12	0.24	0.20	0.959	0.05
						MVNI + deletion	β_1	0.02	0.26	0.22	0.984	0.05
							β_2	-0.04	0.26	0.22	0.986	0.05
						FCS	β_1	-0.02	0.23	0.21	0.976	0.04
β_2	-0.08	0.24					0.21	0.971	0.05			
FCS + deletion	β_1	0.02				0.26	0.22	0.983	0.05			
	β_2	-0.04				0.26	0.22	0.986	0.05			
CCA	β_1	0.00				0.26	0.26	0.942	0.07			
	β_2	0.01				0.24	0.24	0.953	0.06			
0.30	Log(3)	0.30				MVNI	β_1	-0.22	0.17	0.13	0.820	0.07
							β_2	-0.22	0.18	0.14	0.842	0.07
			MVNI + deletion	β_1	-0.06	0.20	0.16	0.983	0.03			
				β_2	-0.09	0.20	0.16	0.970	0.03			
			FCS	β_1	-0.12	0.17	0.14	0.939	0.04			
				β_2	-0.16	0.18	0.15	0.906	0.05			
			FCS + deletion	β_1	-0.06	0.20	0.16	0.981	0.03			
				β_2	-0.10	0.20	0.16	0.968	0.03			
			CCA	β_1	0.00	0.20	0.21	0.947	0.04			
				β_2	-0.00	0.18	0.18	0.956	0.03			
			0.30	Log(3)	0.70	MVNI	β_1	-0.17	0.24	0.19	0.946	0.07
							β_2	-0.25	0.24	0.19	0.887	0.10
MVNI + deletion	β_1	-0.00				0.28	0.22	0.985	0.05			
	β_2	-0.12				0.27	0.22	0.969	0.06			
FCS	β_1	-0.05				0.24	0.21	0.977	0.05			
	β_2	-0.19				0.24	0.21	0.921	0.08			
FCS + deletion	β_1	0.01				0.28	0.22	0.987	0.05			
	β_2	-0.12				0.27	0.22	0.966	0.06			
CCA	β_1	0.01				0.26	0.27	0.947	0.07			
	β_2	0.01				0.24	0.24	0.947	0.06			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.10. Simulation results for X_1 and X_2 continuous, coordinated mechanism, $\lambda = 2$.

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.10	Log(2)	0.30	MVNI	β_1	-0.27	0.31	0.24	0.933	0.13			
				β_2	-0.13	0.33	0.28	0.965	0.09			
			MVNI + deletion	β_1	-0.01	0.38	0.36	0.969	0.13			
				β_2	-0.02	0.35	0.33	0.967	0.11			
			FCS	β_1	-0.03	0.35	0.33	0.963	0.11			
				β_2	-0.04	0.34	0.32	0.964	0.10			
			FCS + deletion	β_1	-0.01	0.38	0.36	0.969	0.13			
				β_2	-0.02	0.35	0.33	0.966	0.11			
			CCA	β_1	-0.01	0.45	0.45	0.945	0.20			
				β_2	-0.01	0.34	0.33	0.947	0.11			
			0.10	Log(2)	0.70	MVNI	β_1	-0.32	0.44	0.35	0.939	0.23
							β_2	-0.13	0.45	0.40	0.969	0.17
MVNI + deletion	β_1	0.01				0.51	0.47	0.964	0.23			
	β_2	-0.01				0.49	0.47	0.959	0.22			
FCS	β_1	-0.02				0.47	0.45	0.964	0.20			
	β_2	-0.05				0.46	0.45	0.961	0.20			
FCS + deletion	β_1	0.01				0.51	0.47	0.963	0.23			
	β_2	-0.01				0.49	0.47	0.962	0.22			
CCA	β_1	0.02				0.57	0.58	0.941	0.34			
	β_2	-0.00				0.47	0.48	0.940	0.23			
0.10	Log(3)	0.30				MVNI	β_1	-0.46	0.32	0.25	0.758	0.27
							β_2	-0.21	0.35	0.31	0.934	0.14
			MVNI + deletion	β_1	0.00	0.42	0.39	0.964	0.15			
				β_2	-0.03	0.38	0.37	0.959	0.14			
			FCS	β_1	-0.07	0.35	0.34	0.953	0.12			
				β_2	-0.11	0.35	0.34	0.952	0.13			
			FCS + deletion	β_1	0.01	0.42	0.39	0.969	0.15			
				β_2	-0.03	0.38	0.37	0.958	0.13			
			CCA	β_1	0.03	0.50	0.51	0.944	0.26			
				β_2	-0.00	0.36	0.38	0.935	0.15			
			0.10	Log(3)	0.70	MVNI	β_1	-0.56	0.48	0.39	0.838	0.47
							β_2	-0.22	0.50	0.43	0.958	0.24
MVNI + deletion	β_1	0.01				0.58	0.55	0.965	0.30			
	β_2	-0.03				0.55	0.53	0.959	0.28			
FCS	β_1	-0.08				0.51	0.49	0.961	0.25			
	β_2	-0.14				0.50	0.48	0.950	0.25			
FCS + deletion	β_1	0.02				0.58	0.55	0.964	0.30			
	β_2	-0.04				0.55	0.53	0.961	0.28			
CCA	β_1	0.01				0.66	0.67	0.943	0.45			
	β_2	0.01				0.53	0.55	0.936	0.30			
0.30	Log(2)	0.30				MVNI	β_1	-0.16	0.17	0.14	0.919	0.04
							β_2	-0.12	0.17	0.14	0.940	0.03

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
			MVNI + deletion	β_1	0.00	0.22	0.18	0.984	0.03			
				β_2	-0.02	0.19	0.17	0.975	0.03			
			FCS	β_1	-0.06	0.18	0.16	0.970	0.03			
				β_2	-0.08	0.18	0.15	0.957	0.03			
			FCS + deletion	β_1	0.00	0.22	0.18	0.985	0.03			
				β_2	-0.02	0.19	0.17	0.978	0.03			
			CCA	β_1	0.00	0.23	0.23	0.947	0.05			
				β_2	0.00	0.17	0.17	0.949	0.03			
0.30	Log(2)	0.70	MVNI	β_1	-0.17	0.24	0.20	0.934	0.07			
				β_2	-0.14	0.24	0.20	0.945	0.06			
			MVNI + deletion	β_1	0.02	0.29	0.25	0.977	0.06			
				β_2	-0.04	0.27	0.24	0.969	0.06			
			FCS	β_1	-0.05	0.24	0.22	0.966	0.05			
				β_2	-0.10	0.24	0.22	0.951	0.06			
			FCS + deletion	β_1	0.02	0.29	0.25	0.978	0.06			
				β_2	-0.04	0.27	0.24	0.968	0.06			
			CCA	β_1	-0.00	0.30	0.31	0.941	0.09			
				β_2	0.00	0.24	0.25	0.940	0.06			
			0.30	Log(3)	0.30	MVNI	β_1	-0.29	0.18	0.14	0.689	0.11
							β_2	-0.24	0.18	0.15	0.789	0.08
MVNI + deletion	β_1	-0.04				0.24	0.19	0.984	0.04			
	β_2	-0.08				0.20	0.17	0.970	0.03			
FCS	β_1	-0.16				0.18	0.15	0.909	0.05			
	β_2	-0.19				0.18	0.15	0.849	0.06			
FCS + deletion	β_1	-0.04				0.24	0.19	0.983	0.04			
	β_2	-0.08				0.20	0.17	0.969	0.03			
CCA	β_1	0.01				0.25	0.25	0.951	0.06			
	β_2	0.01				0.18	0.18	0.948	0.03			
0.30	Log(3)	0.70				MVNI	β_1	-0.26	0.26	0.20	0.890	0.11
							β_2	-0.27	0.25	0.20	0.859	0.11
			MVNI + deletion	β_1	0.01	0.31	0.26	0.978	0.07			
				β_2	-0.11	0.28	0.23	0.963	0.07			
			FCS	β_1	-0.09	0.26	0.22	0.962	0.06			
				β_2	-0.24	0.24	0.21	0.878	0.10			
			FCS + deletion	β_1	0.02	0.31	0.26	0.980	0.07			
				β_2	-0.12	0.28	0.23	0.963	0.07			
			CCA	β_1	0.02	0.32	0.32	0.951	0.11			
				β_2	-0.00	0.25	0.26	0.950	0.07			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.11. Simulation results for X_1 and X_2 continuous, opposite mechanism, $\lambda = 1$.

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.10	Log(2)	0.30	MVNI	β_1	-0.18	0.29	0.24	0.948	0.09			
				β_2	-0.09	0.32	0.29	0.962	0.09			
			MVNI + deletion	β_1	-0.01	0.33	0.31	0.961	0.10			
				β_2	0.01	0.35	0.34	0.960	0.11			
			FCS	β_1	-0.02	0.31	0.30	0.956	0.09			
				β_2	-0.01	0.34	0.33	0.962	0.11			
			FCS + deletion	β_1	-0.01	0.33	0.31	0.961	0.10			
				β_2	0.01	0.35	0.34	0.962	0.11			
			CCA	β_1	0.00	0.37	0.39	0.936	0.15			
				β_2	0.01	0.33	0.33	0.944	0.11			
			0.10	Log(2)	0.70	MVNI	β_1	-0.20	0.41	0.35	0.961	0.16
							β_2	-0.10	0.43	0.39	0.969	0.16
MVNI + deletion	β_1	0.00				0.46	0.43	0.966	0.18			
	β_2	0.00				0.47	0.45	0.961	0.21			
FCS	β_1	-0.01				0.43	0.41	0.964	0.17			
	β_2	-0.02				0.45	0.44	0.961	0.20			
FCS + deletion	β_1	0.01				0.46	0.43	0.965	0.18			
	β_2	-0.00				0.47	0.45	0.960	0.21			
CCA	β_1	0.00				0.47	0.48	0.946	0.23			
	β_2	-0.00				0.43	0.45	0.943	0.20			
0.10	Log(3)	0.30				MVNI	β_1	-0.29	0.30	0.23	0.896	0.14
							β_2	-0.15	0.33	0.28	0.958	0.10
			MVNI + deletion	β_1	-0.01	0.34	0.31	0.970	0.10			
				β_2	0.02	0.36	0.33	0.966	0.11			
			FCS	β_1	-0.04	0.31	0.29	0.963	0.09			
				β_2	-0.03	0.34	0.31	0.965	0.10			
			FCS + deletion	β_1	-0.00	0.34	0.31	0.973	0.10			
				β_2	0.01	0.36	0.32	0.966	0.11			
			CCA	β_1	-0.00	0.37	0.38	0.947	0.14			
				β_2	0.01	0.32	0.32	0.949	0.10			
			0.10	Log(3)	0.70	MVNI	β_1	-0.34	0.42	0.35	0.922	0.24
							β_2	-0.16	0.45	0.38	0.972	0.17
MVNI + deletion	β_1	-0.01				0.48	0.44	0.972	0.19			
	β_2	0.02				0.49	0.45	0.969	0.21			
FCS	β_1	-0.05				0.44	0.41	0.961	0.17			
	β_2	-0.03				0.46	0.43	0.964	0.19			
FCS + deletion	β_1	-0.01				0.48	0.44	0.968	0.19			
	β_2	0.02				0.49	0.45	0.970	0.21			
CCA	β_1	-0.01				0.48	0.48	0.941	0.23			
	β_2	0.01				0.44	0.45	0.946	0.20			
0.30	Log(2)	0.30				MVNI	β_1	-0.10	0.16	0.13	0.952	0.03
							β_2	-0.06	0.18	0.15	0.975	0.03
			MVNI + deletion	β_1	-0.00	0.19	0.16	0.983	0.02			

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
			FCS	β_2	0.02	0.20	0.17	0.976	0.03			
				β_1	-0.03	0.16	0.15	0.971	0.02			
			FCS + deletion	β_2	-0.01	0.18	0.16	0.970	0.03			
				β_1	0.00	0.19	0.16	0.984	0.02			
			CCA	β_2	0.02	0.20	0.17	0.976	0.03			
				β_1	0.00	0.19	0.19	0.941	0.04			
<hr/>												
0.30	Log(2)	0.70	MVNI	β_1	-0.12	0.23	0.19	0.960	0.05			
				β_2	-0.06	0.24	0.20	0.977	0.05			
			MVNI + deletion	β_1	-0.01	0.26	0.22	0.979	0.05			
				β_2	0.02	0.27	0.23	0.974	0.05			
			FCS	β_1	-0.04	0.23	0.20	0.970	0.04			
				β_2	-0.01	0.24	0.22	0.972	0.05			
			FCS + deletion	β_1	-0.01	0.26	0.22	0.981	0.05			
				β_2	0.02	0.27	0.23	0.976	0.05			
			CCA	β_1	0.01	0.24	0.24	0.948	0.06			
				β_2	-0.00	0.22	0.22	0.945	0.05			
			<hr/>									
			0.30	Log(3)	0.30	MVNI	β_1	-0.14	0.17	0.13	0.940	0.04
β_2	-0.10	0.19					0.15	0.958	0.03			
MVNI + deletion	β_1	0.02				0.20	0.16	0.989	0.03			
	β_2	0.02				0.21	0.17	0.981	0.03			
FCS	β_1	-0.02				0.17	0.14	0.987	0.02			
	β_2	-0.03				0.19	0.16	0.979	0.03			
FCS + deletion	β_1	0.02				0.20	0.16	0.989	0.03			
	β_2	0.01				0.21	0.17	0.986	0.03			
CCA	β_1	0.01				0.19	0.19	0.943	0.04			
	β_2	0.01				0.17	0.17	0.951	0.03			
<hr/>												
0.30	Log(3)	0.70				MVNI	β_1	-0.17	0.23	0.18	0.946	0.06
			β_2	-0.11	0.25		0.20	0.956	0.05			
			MVNI + deletion	β_1	0.01	0.26	0.21	0.983	0.04			
				β_2	0.02	0.27	0.23	0.979	0.05			
			FCS	β_1	-0.04	0.23	0.19	0.981	0.04			
				β_2	-0.04	0.25	0.22	0.970	0.05			
			FCS + deletion	β_1	0.01	0.26	0.21	0.985	0.04			
				β_2	0.01	0.27	0.23	0.979	0.05			
			CCA	β_1	0.00	0.23	0.24	0.940	0.06			
				β_2	0.00	0.22	0.23	0.933	0.05			

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Table 5.12. Simulation results for X_1 and X_2 continuous, opposite mechanism, $\lambda = 2$.

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
0.10	Log(2)	0.30	MVNI	β_1	-0.27	0.31	0.24	0.932	0.13			
				β_2	-0.09	0.34	0.32	0.961	0.11			
			MVNI + deletion	β_1	-0.00	0.38	0.36	0.966	0.13			
				β_2	0.02	0.37	0.37	0.948	0.14			
			FCS	β_1	-0.02	0.35	0.33	0.959	0.11			
				β_2	-0.01	0.35	0.36	0.946	0.13			
			FCS + deletion	β_1	0.00	0.38	0.36	0.965	0.13			
				β_2	0.02	0.37	0.37	0.946	0.14			
			CCA	β_1	0.01	0.49	0.49	0.939	0.24			
				β_2	0.01	0.34	0.37	0.931	0.14			
			0.10	Log(2)	0.70	MVNI	β_1	-0.31	0.44	0.36	0.939	0.23
							β_2	-0.12	0.46	0.41	0.963	0.18
MVNI + deletion	β_1	0.03				0.52	0.49	0.962	0.24			
	β_2	-0.01				0.50	0.48	0.956	0.24			
FCS	β_1	-0.00				0.47	0.45	0.960	0.21			
	β_2	-0.05				0.48	0.46	0.957	0.21			
FCS + deletion	β_1	0.04				0.52	0.49	0.965	0.24			
	β_2	-0.01				0.50	0.48	0.962	0.23			
CCA	β_1	0.03				0.59	0.59	0.942	0.35			
	β_2	-0.02				0.46	0.48	0.939	0.23			
0.10	Log(3)	0.30				MVNI	β_1	-0.45	0.32	0.25	0.783	0.27
							β_2	-0.18	0.35	0.30	0.952	0.12
			MVNI + deletion	β_1	0.01	0.41	0.39	0.967	0.15			
				β_2	0.01	0.38	0.36	0.959	0.13			
			FCS	β_1	-0.05	0.35	0.34	0.953	0.12			
				β_2	-0.06	0.36	0.34	0.951	0.12			
			FCS + deletion	β_1	0.02	0.41	0.39	0.966	0.15			
				β_2	0.01	0.38	0.36	0.962	0.13			
			CCA	β_1	0.01	0.50	0.51	0.946	0.26			
				β_2	-0.00	0.35	0.35	0.938	0.13			
			0.10	Log(3)	0.70	MVNI	β_1	-0.58	0.47	0.37	0.830	0.47
							β_2	-0.17	0.48	0.42	0.961	0.21
MVNI + deletion	β_1	-0.00				0.56	0.52	0.966	0.28			
	β_2	0.02				0.53	0.51	0.959	0.26			
FCS	β_1	-0.08				0.48	0.46	0.961	0.22			
	β_2	-0.07				0.49	0.47	0.959	0.22			
FCS + deletion	β_1	0.01				0.56	0.52	0.971	0.27			
	β_2	0.01				0.53	0.51	0.962	0.26			
CCA	β_1	0.00				0.60	0.62	0.939	0.39			
	β_2	0.01				0.48	0.50	0.938	0.25			
0.30	Log(2)	0.30				MVNI	β_1	-0.15	0.17	0.14	0.924	0.04
							β_2	-0.07	0.19	0.16	0.963	0.03
			MVNI + deletion	β_1	0.01	0.22	0.18	0.981	0.03			

Outcome prevalence	β_1, β_2	$\text{Corr}(X_1, X_2)$	Method	Parameter	Bias	Avg SE	Emp SE	Coverage	MSE			
			FCS	β_2	0.02	0.21	0.18	0.972	0.03			
				β_1	-0.05	0.18	0.16	0.969	0.03			
			FCS + deletion	β_2	-0.03	0.19	0.17	0.968	0.03			
				β_1	0.01	0.22	0.19	0.980	0.03			
			CCA	β_2	0.02	0.21	0.18	0.971	0.03			
				β_1	0.01	0.25	0.24	0.954	0.06			
0.30	Log(2)	0.70	MVNI	β_2	-0.01	0.17	0.18	0.942	0.03			
				β_1	-0.19	0.24	0.19	0.933	0.07			
			MVNI + deletion	β_2	-0.07	0.25	0.21	0.972	0.05			
				β_1	-0.01	0.29	0.24	0.981	0.06			
			FCS	β_2	0.03	0.28	0.24	0.981	0.06			
				β_1	-0.08	0.24	0.21	0.966	0.05			
			FCS + deletion	β_2	-0.03	0.25	0.22	0.970	0.05			
				β_1	-0.01	0.29	0.24	0.982	0.06			
			CCA	β_2	0.03	0.28	0.24	0.978	0.06			
				β_1	0.00	0.29	0.29	0.945	0.09			
			0.30	Log(3)	0.30	MVNI	β_2	-0.00	0.23	0.23	0.954	0.05
							β_1	-0.18	0.18	0.13	0.912	0.05
MVNI + deletion	β_2	-0.09				0.20	0.16	0.962	0.03			
	β_1	0.04				0.23	0.18	0.983	0.04			
FCS	β_2	0.05				0.22	0.18	0.977	0.04			
	β_1	-0.03				0.18	0.15	0.975	0.02			
FCS + deletion	β_2	-0.03				0.20	0.17	0.972	0.03			
	β_1	0.05				0.23	0.18	0.982	0.04			
CCA	β_2	0.04				0.22	0.18	0.978	0.04			
	β_1	0.01				0.24	0.25	0.946	0.06			
0.30	Log(3)	0.70				MVNI	β_2	0.01	0.17	0.18	0.945	0.03
							β_1	-0.25	0.24	0.19	0.886	0.10
			MVNI + deletion	β_2	-0.07	0.26	0.20	0.981	0.05			
				β_1	0.01	0.30	0.25	0.983	0.06			
			FCS	β_2	0.06	0.29	0.23	0.980	0.06			
				β_1	-0.07	0.24	0.21	0.974	0.05			
			FCS + deletion	β_2	-0.02	0.26	0.22	0.980	0.05			
				β_1	0.02	0.29	0.25	0.983	0.06			
			CCA	β_2	0.05	0.29	0.23	0.982	0.06			
				β_1	-0.00	0.29	0.29	0.945	0.08			
							β_2	0.01	0.23	0.22	0.949	0.05

Abbreviations: MVNI, multivariate normal imputation; FCS, fully conditional specification; CCA, complete case analysis; Avg SE, average standard error; Emp SE, empirical standard error; MSE, mean square error.

Web appendix C. Full results from null-case sensitivity analyses.

Table 5.13. Bias in scenarios with X_1 and X_2 binary, coordinated missing data mechanism.

Simulation scenario	Parameter	MVNI	MVNI + deletion	FCS	FCS + deletion
1. Outcome prevalence = 0.10, $RR(X_1, X_2) = 3$, $\beta_1 = \beta_2 = \log(3)$, $\lambda = 2$	β_1	-0.03	0.10	0.05	0.04
	β_2	-0.28	-0.10	0.00	0.02
2. As in (1.), but with $RR(X_1, X_2) = 1$	β_1	-0.14	0.01	0.01	0.01
	β_2	-0.28	-0.08	0.01	0.03
3. As in (1.), but with $\beta_1 = 0$	β_1	0.02	0.05	0.01	0.01
	β_2	-0.17	-0.05	0.00	0.01
4. As in (1.), but with $\beta_2 = 0$	β_1	-0.15	0.00	0.00	0.00
	β_2	0.02	0.02	0.01	0.01
5. As in (1.), but with $\beta_1 = \beta_2 = 0$	β_1	0.00	0.00	0.00	0.00
	β_2	0.00	-0.01	-0.02	-0.01
6. As in (1.), but with $\lambda = 0$ (MCAR)	β_1	-0.12	0.11	0.04	0.04
	β_2	-0.28	-0.10	0.05	0.06
7. Outcome prevalence = 0.30, $RR(X_1, X_2) = 3$, $\beta_1 = \beta_2 = \log(3)$, $\lambda = 2$	β_1	0.07	0.07	0.04	0.02
	β_2	-0.31	-0.14	-0.10	-0.04
8. As in (7.), but with $RR(X_1, X_2) = 1$	β_1	-0.07	-0.01	-0.01	-0.01
	β_2	-0.32	-0.16	-0.12	-0.07
9. As in (7.), but with $\beta_1 = 0$	β_1	0.03	0.03	0.00	0.00
	β_2	-0.12	-0.05	0.01	0.01
10. As in (7.), but with $\beta_2 = 0$	β_1	-0.06	0.00	0.00	0.00
	β_2	0.00	0.00	0.00	0.00
11. As in (7.), but with $\beta_1 = \beta_2 = 0$	β_1	0.00	0.00	0.00	0.00
	β_2	0.00	0.00	0.00	0.00
12. As in (7.), but with $\lambda = 0$ (MCAR)	β_1	-0.04	0.07	0.00	0.01
	β_2	-0.21	-0.11	0.00	0.01

Abbreviations: RR, relative risk; MCAR, missing completely at random; MVNI, multivariate normal imputation; FCS, fully conditional specification.

Table 5.14. Bias in scenarios with X_1 and X_2 continuous, coordinated missing data mechanism.

Simulation scenario	Parameter	MVNI	MVNI + deletion	FCS	FCS + deletion
1. Outcome prevalence = 0.10, $\text{Corr}(X_1, X_2) = 0.70$, $\beta_1 = \beta_2 = \log(3)$, $\lambda = 2$	β_1	-0.56	0.01	-0.08	0.02
	β_2	-0.22	-0.03	-0.14	-0.04
2. As in (1.), but with $\text{Corr}(X_1, X_2) = 0$	β_1	-0.38	0.00	-0.06	0.00
	β_2	-0.20	-0.02	-0.08	-0.02
3. As in (1.), but with $\beta_1 = 0$	β_1	-0.13	0.01	0.01	0.01
	β_2	-0.17	0.00	-0.04	0.00
4. As in (1.), but with $\beta_2 = 0$	β_1	-0.37	0.01	-0.04	0.01
	β_2	0.00	0.00	0.00	-0.01
5. As in (1.), but with $\beta_1 = \beta_2 = 0$	β_1	0.00	0.00	0.01	0.00
	β_2	0.00	0.00	0.00	0.00
6. As in (1.), but with $\lambda = 0$ (MCAR)	β_1	-0.15	0.00	0.00	0.00
	β_2	-0.17	-0.03	-0.03	-0.03
7. Outcome prevalence = 0.30, $\text{Corr}(X_1, X_2) = 0.70$, $\beta_1 = \beta_2 = \log(3)$, $\lambda = 2$	β_1	-0.26	0.01	-0.09	0.02
	β_2	-0.27	-0.11	-0.24	-0.12
8. As in (7.), but with $\text{Corr}(X_1, X_2) = 0$	β_1	-0.27	-0.05	-0.15	-0.05
	β_2	-0.21	-0.06	-0.16	-0.06
9. As in (7.), but with $\beta_1 = 0$	β_1	-0.04	0.02	0.01	0.02
	β_2	-0.17	-0.03	-0.10	-0.03
10. As in (7.), but with $\beta_2 = 0$	β_1	-0.24	0.00	-0.10	0.00
	β_2	0.00	0.00	0.00	0.00
11. As in (7.), but with $\beta_1 = \beta_2 = 0$	β_1	-0.01	-0.01	-0.01	-0.01
	β_2	0.01	0.01	0.01	0.01
12. As in (7.), but with $\lambda = 0$ (MCAR)	β_1	-0.11	0.00	0.00	0.00
	β_2	-0.17	-0.08	-0.08	-0.08

Abbreviations: Corr, correlation; MCAR, missing completely at random; MVNI, multivariate normal imputation; FCS, fully conditional specification.

** End of submitted article ***

5.4. Additional discussion

Since the data generation models in the simulation study were known, the shortcomings of MI could be attributed to differences between the imputation and data generation models (i.e. imputation model misspecification). Alternatively, since the analysis models were equivalent to the data generation models, the shortcomings of MI could be attributed to inconsistencies between imputation and analysis models (i.e. imputation model incompatibility). In practice the log

binomial analysis model could also be misspecified, in which case performance deficits due to imputation model misspecification might differ to those of imputation model incompatibility. Which is the larger concern in practice is beyond the scope of this thesis. It should be reiterated that if the log binomial analysis model is deemed inappropriate for a given dataset due to apparent misspecification, relative risks can instead be estimated from a model with a different link function (e.g. logistic regression) by applying marginal or conditional standardisation to predicted probabilities (64).

6. Multiple imputation in randomised trials

6.1. Preface

This chapter presents the third article contributing to this thesis, published in *Statistical Methods in Medical Research*. The aims of the article are to evaluate the performance of MI for handling missing outcome data in RCTs and to explore the merits of imputing overall and separately by randomised group in this context. The article covers the common scenarios of missing data in a continuous or a binary outcome variable measured once or repeatedly over time, where interest lies in estimating the effect of treatment according to the ITT principle. The article also considers the use of MI for handling missing data in a baseline covariate for adjustment, for example a baseline measure of the outcome variable. Although not a missing outcome data problem, which is the focus of this thesis, the use of MI for handling missing data in a baseline covariate is considered in the article for the sake of completeness (as another common scenario encountered in the analysis of RCTs). Relevant literature on handling missing baseline data in RCTs is described within the article.

6.2. Statement of authorship

Title of paper	Should multiple imputation be the method of choice for handling missing data in randomized trials?
Publication status	Accepted for publication
Publication details	Should multiple imputation be the method of choice for handling missing data in randomized trials? <i>Statistical Methods in Medical Research</i> , 2016; electronic publication available on journal website ahead of print, doi 10.1177/0962280216683570.

Principal author

Name (Candidate)	Thomas Sullivan		
Contribution	Designed the study, simulated the data, performed all analyses, interpreted the results, drafted the manuscript and acted as corresponding author.		
Overall percentage (%)	75		
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	11/07/2017

Co-author contributions

By signing the Statement of authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and

- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of co-author	Ian White		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	14/7/2017

Name of co-author	Amy Salter		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	11/07/2017

Name of co-author	Phillip Ryan		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	12/07/2017

Name of co-author	Katherine Lee		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	26/07/2017

6.3. Article

In this section, I provide the text, tables, figures, and appendices from the published manuscript.

6.3.1. Abstract

The use of multiple imputation has increased markedly in recent years, and journal reviewers may expect to see multiple imputation used to handle missing data. However in randomised trials, where treatment group is always observed and independent of baseline covariates, other approaches may be preferable. Using data simulation we evaluated multiple imputation, performed both overall and separately by randomised group, across a range of commonly encountered scenarios. We considered both missing outcome and missing baseline data, with missing outcome data induced under missing at random mechanisms. Provided the analysis model was correctly specified, multiple imputation produced unbiased treatment effect estimates, but alternative unbiased approaches were often more efficient. When the analysis model overlooked an interaction effect involving randomised group, multiple imputation produced biased estimates of the average treatment effect when applied to missing outcome data, unless imputation was performed separately by randomised group. Based on these results, we conclude that multiple imputation should not be seen as the only acceptable way to handle missing data in randomised trials. In settings where multiple imputation is adopted, we recommend that imputation is carried out separately by randomised group.

6.3.2. Introduction

Research articles and guidance documents have emphasised the role of prevention in minimising the impact of missing data (65, 71, 107, 108), but most randomised controlled trials (RCTs) have some missing data (109). Given the potential for

biased and inefficient treatment effect estimates, it is crucial that missing data are handled appropriately during the analysis.

All statistical analyses involve assumptions about the mechanism responsible for the missing data. Rubin (4) introduced three classes of mechanisms for missing data: missing completely at random (MCAR), where the probability of missingness is unrelated to observed or unobserved data; missing at random (MAR), where the probability of missingness is unrelated to unobserved data conditional on observed data; and missing not at random (MNAR), where the probability of missingness depends on unobserved data conditional on observed data. Since MAR and MNAR cannot be distinguished from observed data, it is essential that the assumptions of the analytic approach are scientifically plausible and clearly stated (1, 20). To assess the robustness of findings to the assumption made about the missing data mechanism in the primary analysis of an RCT, additional sensitivity analyses are strongly recommended (1, 13, 19, 20, 110).

Multiple imputation (MI) (4) is a statistical approach to handling missing data that has been widely adopted due to its flexibility and ease of implementation (17, 29). MI involves fitting a statistical model to the observed data and using it to estimate values for the missing data. To incorporate missing data uncertainty, multiple values are imputed for each missing observation, producing multiple complete datasets. Following analysis of these datasets using standard complete data techniques, the multiple parameter estimates are combined using Rubin's rules (4) to give a single MI estimate. Standard implementations of MI assume that data are MAR, although it can also be applied under an MNAR assumption (4). Provided the assumption about the missing data mechanism is met and models used for imputation and analysis are correctly specified, MI produces consistent and asymptotically efficient parameter estimates with nominal coverage (4). Of the various methods of imputation available, MI based on the multivariate normal distribution (11) and MI by chained equations (5, 9, 10) are most commonly used in RCTs (7).

With the use of MI in RCTs rising dramatically in recent years (7, 17, 111), editors and journal reviewers may expect to see MI used to handle missing data. Indeed, we are aware of several recent instances where reviewers have pushed with little justification for trial data to be re-analysed using MI. However, whether MI should be viewed as the gold standard approach for handling missing data in RCTs is questionable. Importantly, results derived in general regression settings supporting the use of MI may not be applicable to RCTs. Unlike observational studies, the key exposure in RCTs (randomised group) is always observed and known to be independent of baseline covariates. In addition, missing data occur primarily in the outcome variable, although baseline covariates may also have missing data. Under these conditions, some of the value of MI may be lost and other methods of analysis may be preferable.

Another uncertainty around the use of MI in RCTs is whether imputation is best carried out across all participants or separately by randomised group. If subgroup analyses are of interest, it is essential that interaction terms are accounted for in the imputation process to avoid biasing interaction tests towards the null. Rather than specifying interaction terms within the imputation model, several authors have recommended fitting separate imputation models within each randomised group (13-15, 102). This strategy is appealing due to its simplicity and ability to facilitate subgroup comparisons for any baseline covariate included in the imputation model. Unfortunately its performance is not well understood, and it is unclear how imputation should proceed when subgroup analyses are not of interest and the intention is to only produce average treatment effects from main effects models.

This article describes the performance of MI in the RCT setting, covering the common scenarios of missing data in an outcome measured once or repeatedly over time and missing data in a baseline covariate. Using a series of illustrative data simulations and a case study, we compare MI with other standard approaches for handling missing data and explore the merits of imputing overall and separately by randomised group. Throughout we assume that missing data are

unplanned rather than by design, and that interest lies in estimating the effect of treatment according to the intention to treat (ITT) principle. If treatment discontinuations occur, we therefore assume the aim is to estimate a “de facto” estimand (66, 112) and that data are equally available before and after treatment discontinuations; we consider the case where data cannot be collected after treatment discontinuation in the discussion (Section 6.3.9). For missing outcome data, we restrict attention to settings where they are assumed to be MAR, since this assumption is often made in the primary analysis of an RCT and corresponds with the standard implementation of MI.

The remainder of the article is structured as follows. Section 6.3.3 describes issues in adhering to the ITT principle in the presence of missing data and implications for the use of MI in RCTs. Section 6.3.4 defines key notation and outlines general simulation methods for evaluating the performance of MI. Section 6.3.5 focuses on the performance of MI for handling missing data in an outcome measured at a single time point. Section 6.3.6 considers missing data in an outcome measured repeatedly over time and the use of auxiliary variables in MI, while Section 6.3.7 focuses on missing data in a baseline covariate for adjustment. Section 6.3.8 shows the application of MI to the DINO trial. Finally, conclusions and general recommendations are provided in Section 6.3.9.

6.3.3. Intention to treat and missing data

The goal of ITT, or analysing as randomised, is to maintain the balance in prognostic factors achieved by randomisation, which is critical for avoiding selection bias and establishing causation (69, 70). In addition to preserving the benefits of randomisation, an ITT analysis may better inform changes in subsequent clinical practice, where patients do not always comply with treatment. Following the ITT principle entails estimating the ITT estimand, which is defined as the average effect of randomisation, irrespective of treatment received, over all randomised individuals (68). Due to fluctuating use of the term ITT, this has more recently been described as a de facto estimand (66, 112). Interest in the ITT

estimand has implications both for trial conduct and analysis. First, attempts should be made to collect outcome data on all randomised participants, irrespective of adherence to the protocol (1, 20, 68). For example, outcome data should still be retrieved for participants that discontinue or switch treatments during the course of a trial. Second, all collected outcome data should be included in the analysis, including data from participants that deviate from the protocol (1, 20, 68). Although there are settings where it may not be feasible to measure outcomes following a protocol deviation, or where exclusion of collected outcome data may be justifiable, we do not tackle these scenarios in this article.

Despite efforts to collect data on all randomised participants, invariably there will be some missing data. Exactly what constitutes an ITT analysis in the presence of missing data has been much debated (113). Some researchers have suggested that missing outcome data ought to be imputed, so that the full randomised sample can be included in the analysis (70, 72, 73). Others have argued that imputation is unnecessary and that an ITT analysis need only provide a valid estimate of the ITT estimand (1, 20, 74). Given recent commentary on the importance of defining and validly estimating the causal estimand of interest (1), and noting that none of the current guidance documents strictly recommend imputing missing outcomes, we adopt the second view. In differentiating between competing statistical methods, we therefore focus on their capacity to provide an unbiased and precise estimate of the ITT estimand rather than their ability to include all randomised participants.

6.3.4. Methods

Setting

Let Y_i and X_i define values for the i^{th} participant ($i = 1$ to n) on an outcome variable and a baseline variable, respectively. Assume the i^{th} participant is randomised independently to treatment group T_i ($0 = \text{control}$, $1 = \text{new treatment}$) with probability 0.5. Let M_{Y_i} and M_{X_i} denote whether Y_i and X_i are missing or

observed (1 = missing, 0 = observed). In the absence of missing data, suppose the adjusted analysis model

$$g(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 X_i \quad (1)$$

is of interest, where $\mu_i = E(Y_i|T_i, X_i)$ and g is an appropriate link function. Of principal importance is the (adjusted) treatment coefficient β_1 . Note we focus primarily on adjusted estimates in this article, since adjustment for pre-specified baseline covariates is common and can lead to substantial increases in power for testing the effect of treatment (114, 115). As conclusions about treatment are typically based on main effects models (65), we also restrict attention to analysis models that do not include interaction terms.

Multiple imputation

In the first stage of MI, multiple values ($m > 1$) for each missing observation are independently simulated from an imputation model. For missing data restricted to the outcome, the imputation model would typically regress observed values of Y on X and T . Additional auxiliary variables that are not in the analysis model can also be added to the imputation model to improve the prediction of missing values. Let $\hat{\gamma}$ denote the parameter estimates from the imputation model and γ_j^* ($j= 1$ to m) random draws from the posterior distribution of γ . For each random draw, missing values in Y are replaced by simulated values from the posterior predictive distribution of Y according to γ_j^* . For missing data restricted to a baseline covariate, the imputation model instead describes the conditional distribution of X according to Y and T . If MI is performed separately by randomised group, T is omitted from the separate imputation models.

In the second stage of MI, the intended analysis is performed on each of the m complete datasets, in this case model (1). Let $\hat{\theta}_j$ denote the estimate of β_1 from the j^{th} imputed dataset and W_j the corresponding variance estimate. Using Rubin's rules (4), the combined MI treatment effect estimate $\hat{\theta}$ is calculated as the mean of

the m estimates, i.e. $\hat{\theta} = 1/m \sum_{j=1}^m \hat{\theta}_j$. The variance is given by $\text{var}(\hat{\theta}) = W + B(1 + 1/m)$, where $W = 1/m \sum_{j=1}^m W_j$ is the average within-imputation variance and $B = (m - 1)^{-1} \sum_{j=1}^m (\hat{\theta}_j - \hat{\theta})^2$ the between imputation variance. Hypothesis tests and confidence intervals can be obtained using a t -distribution with $\nu = (m - 1)[1 + W/(1 + m^{-1})B]^2$ degrees of freedom.

General simulation methods

Simulation studies were undertaken to describe the performance of MI for handling missing data in a univariate outcome (Section 6.3.5), a multivariate outcome (Section 6.3.6), and a baseline covariate (Section 6.3.7). For each scenario, 2,000 datasets of size $n = 600$ were generated, with 300 observations allocated to each group. The sample size was chosen to be similar to that of a case study (see Section 6.3.8) and to represent a medium-sized trial. Three statistical methods were considered across all settings based on the adjusted analysis model (1): complete case analysis (CCA), MI performed overall, and MI performed by randomised group. For MI, linear and logistic regression were used for the imputation of continuous and binary variables, respectively, with $m = 50$ imputations based on the rule of thumb that the number of imputations should at least equal the percentage of missing data (5). Completed datasets were analysed using linear and logistic regression as appropriate, with treatment effect estimates combined using Rubin's rules (4). Performance was evaluated in terms of bias, empirical standard error (SE), power, and the coverage of estimated 95% confidence intervals. Based on 2,000 simulated datasets, on 95% of occasions the coverage is expected to lie between 0.94 and 0.96 for a true coverage of 0.95. All analyses were performed in SAS version 9.3 (SAS Institute, Inc., Cary, North Carolina).

6.3.5. Missing data in a univariate outcome

When a univariate (once-measured) outcome is MAR conditional on fully observed covariates, a correctly specified CCA with covariate adjustment

produces unbiased and efficient estimates of regression parameters (18, 23, 24). It has also been shown that MI with a large number of imputations approximates a CCA in this setting, provided that imputation and analysis models are the same (81). Using data simulation, we verify these results for RCTs, explore the implications of imputing overall or by randomised group and investigate settings where the analysis model is misspecified.

Correctly specified analysis model

Data were simulated from the model $Y_i = 0.30T_i + \beta_2 X_i + e_i$, with X and $e \sim N(0,1)$. To assess whether model performance depended on the strength of association between X and Y , β_2 was varied so that $\text{corr}(X, Y|T) = 0.30$ or 0.70 . Since comparisons were insensitive to the treatment effect, β_1 was fixed at 0.30 to reflect a small effect size. Following generation of complete datasets, values in Y were set to missing according to three MAR mechanisms:

- 1) MAR X: Odds of missing Y increase by a factor λ per standard deviation (SD) increase in X .
- 2) MAR X+T: Odds of missing Y are λ times higher in the control group and increase by a factor λ per SD increase in X .
- 3) MAR X×T: Odds of missing Y are λ times higher for treatment group participants with $X \leq 0$ and for control group participants with $X > 0$.

Each missingness mechanism was simulated using a logistic regression model, with $\lambda = 1.5$ or 2.5 to indicate weak and strong mechanisms, respectively, and with the model intercept varied to produce 20% (realistic) or 50% (extreme) missing data. This resulted in 24 simulation scenarios (12 missing data scenarios and two values for β_2). Supposing that X is a measure of disease severity, the MAR X and MAR X+T mechanisms might reflect settings where participants with more severe disease or randomised to the control group are more likely to

have missing outcome data. The MAR $X \times T$ mechanism could apply in settings where treatment group participants with less severe disease are also more likely to have missing outcome data due to a perceived lack of need to continue treatment.

As expected, CCA, MI overall and MI by group all produced unbiased treatment effect estimates across the 24 simulation scenarios, with coverage probabilities remaining close to 0.95 throughout (range 0.94, 0.96). Compared to CCA, empirical SEs were on average 0.4% and 2.7% larger with MI overall and MI by group, respectively, which translated to an average loss of power of 0.8% for MI overall and 2.6% for MI by group. Figure 6.1 shows the performance of the various approaches in scenarios with 50% missing data, a strong MAR mechanism, and where $\text{corr}(X, Y|T) = 0.70$; these more extreme scenarios were chosen to highlight differences between approaches. Results from an unadjusted CCA are also displayed for comparison. In all figures, note that error bars indicate estimation efficiency (± 1 empirical SE). Unsurprisingly, MI offered no advantages over a CCA across the range of missingness mechanisms. Of note, unadjusted CCA produced biased estimates when the probability of missing data depended on X and T , with coverage dropping to 0.39 under the MAR $X \times T$ mechanism.

Similar results were obtained from a simulation study involving a binary outcome (see web appendix A, Section 6.3.10; also available online at the journal website).

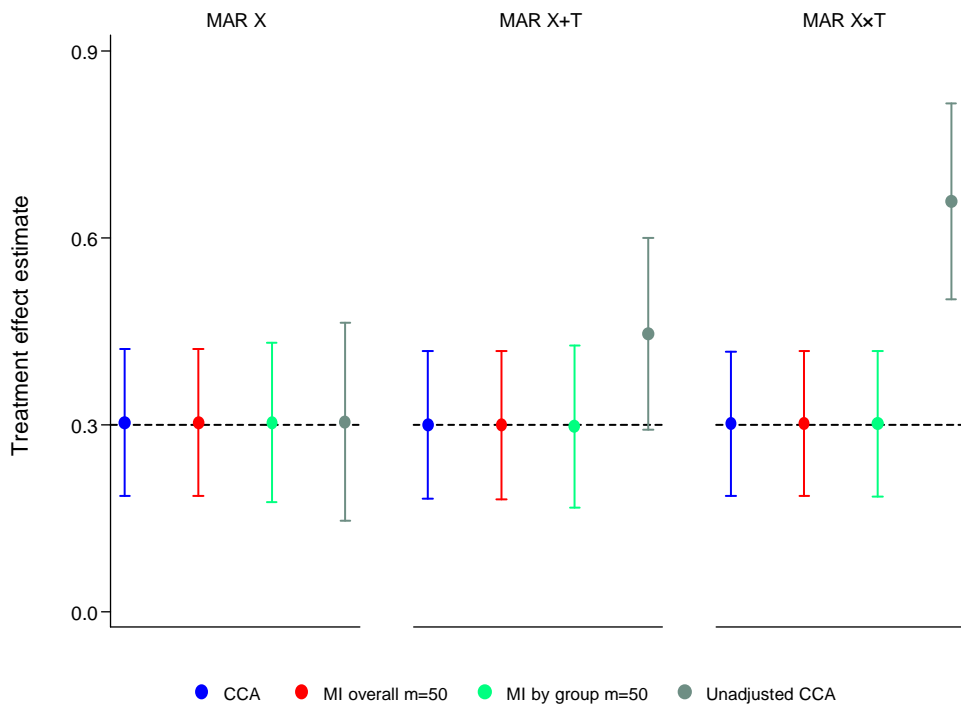


Figure 6.1. Mean treatment effect estimates for 50% missing data in a continuous outcome, $\text{corr}(X, Y|T) = 0.70$, strong MAR mechanisms, correctly specified analysis model. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Misspecified analysis model, continuous outcome

We now consider settings where an interaction between X and T is overlooked in favor of producing an estimate of the average treatment effect (ATE). This approach is common in practice, since ATEs are commonly used to draw conclusions about treatment and are of greater relevance to policy-related questions (65, 116). Further, tests of interaction are often viewed as exploratory and can be underpowered (65). For effect modification by discrete X , we assume that interest lies in estimating the ATE given by $\sum_X \pi_X \alpha_X$, where $\pi_X = P(X = x)$ and α_x denotes the ITT estimand for $X = x$.

Considering binary X with $\pi_0 = \pi_1 = 0.5$, data were generated from the model $Y_i = \beta_1 T_i + 0.30X_i + \beta_3 X_i T_i + e_i$, where $e_i \sim N(0,1)$. Fixing the ATE at 0.30, we investigated both weak ($\beta_1 = \beta_3 = 0.20$ or equivalently $\alpha_0 = 0.20, \alpha_1 = 0.40$) and strong ($\beta_1 = 0, \beta_3 = 0.60$ or $\alpha_0 = 0, \alpha_1 = 0.60$) interaction effects between X and T . Following generation of complete datasets, values in Y were set to missing according to the three mechanisms described earlier. Analysis model (1), misspecified due to the absence of the interaction term between X and T , was the substantive model of interest.

Across all simulation scenarios MI by group produced unbiased estimates of the ATE with nominal coverage (coverage range 0.94, 0.96). In contrast, CCA and MI overall produced biased estimates under the MAR X and MAR $X+T$ mechanisms. Figure 6.2 illustrates performance for the MAR X mechanism in scenarios with 50% missing data. As seen in the figure, the bias of CCA and MI overall increased with the strength of the missing data mechanism and the degree of effect modification. For a strong missing data mechanism and a strong interaction, the ATE was estimated to be 0.17 (absolute bias = 0.13), with coverage dropping to 0.81 for both approaches. Similar results were observed with 20% missing data, although predictably biases were smaller in magnitude (absolute bias ≤ 0.06). Instead of estimating the desired ATE, CCA and MI overall produced an estimate that was weighted by the probability of missing data within strata defined by X and T . In particular, the estimated ATE was proportional to $\sum_X \pi_X \alpha_X R_{0X} R_{1X} / (R_{0X} + R_{1X})$, where $R_{TX} = P(M_Y = 0 | T = t, X = x)$. No bias was observed for these approaches for the MAR $X \times T$ mechanism, since $R_{00} = R_{11}$ and $R_{10} = R_{01}$ under this mechanism. Although the bias of MI overall could be eliminated by including the interaction term in the imputation model (results not shown), this may not be an obvious strategy if subgroup analyses are not of interest.

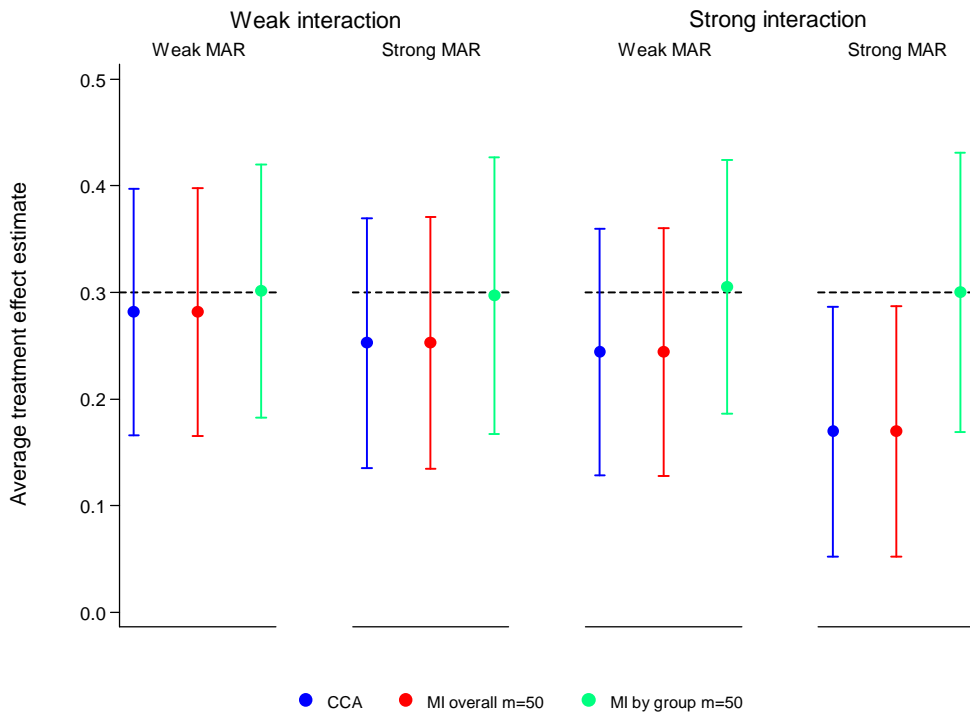


Figure 6.2. Mean average treatment effect estimates for 50% missing data in a continuous outcome under the MAR X mechanism (odds of Y missing 1.5 (weak MAR) or 2.5 (strong MAR) times higher per standard deviation increase in X), incorrectly specified analysis model. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Misspecified analysis model, binary outcome

For binary outcomes, the notion of an ATE from a misspecified logistic regression model is more complex. Assuming effect modification by discrete X, omission of the interaction effect from the analysis model can lead to an ATE estimate that differs substantially from a weighted average of stratum specific effects (on both odds and log odds scales). In this setting, we consider the ATE that would have been observed with complete data as the “least false” ATE. In the presence of missing data, we assume the goal is to reproduce this least false ATE.

Considering binary X with $\pi_0 = \pi_1 = 0.5$, data were generated from the model $\text{logit } P(Y_i = 1) = -1.77 + \beta_1 T_i + 0.69 X_i + \beta_3 X_i T_i$. The intercept value was

chosen so that $P(Y = 1|T = 0) = 0.20$, while the coefficient for X gives $OR(Y, X|T = 0) = 2.0$. Fixing the average of the stratum specific effects on the logit scale at 0.69 ($OR = 2.0$), we evaluated both weak ($\beta_1 = \beta_3 = 0.46$) and strong ($\beta_1 = 0, \beta_3 = 1.38$) interaction effects between X and T . Following generation of complete datasets, values in Y were set to missing according to the three mechanisms described earlier.

Across the 24 simulation scenarios (12 missing data scenarios \times 2 interactions), MI by group was unbiased in reproducing the least false ATE (absolute bias ≤ 0.02), with coverage remaining close to 0.95 (range 0.94, 0.96). In contrast, CCA and MI overall produced biased estimates under the MAR X and MAR X+T mechanisms. Figure 6.3 summarises performance under the MAR X mechanism for 50% missing data. In parallel with results for continuous outcome data, the bias of CCA and MI overall increased with the strength of the missing data mechanism and the interaction between X and T .

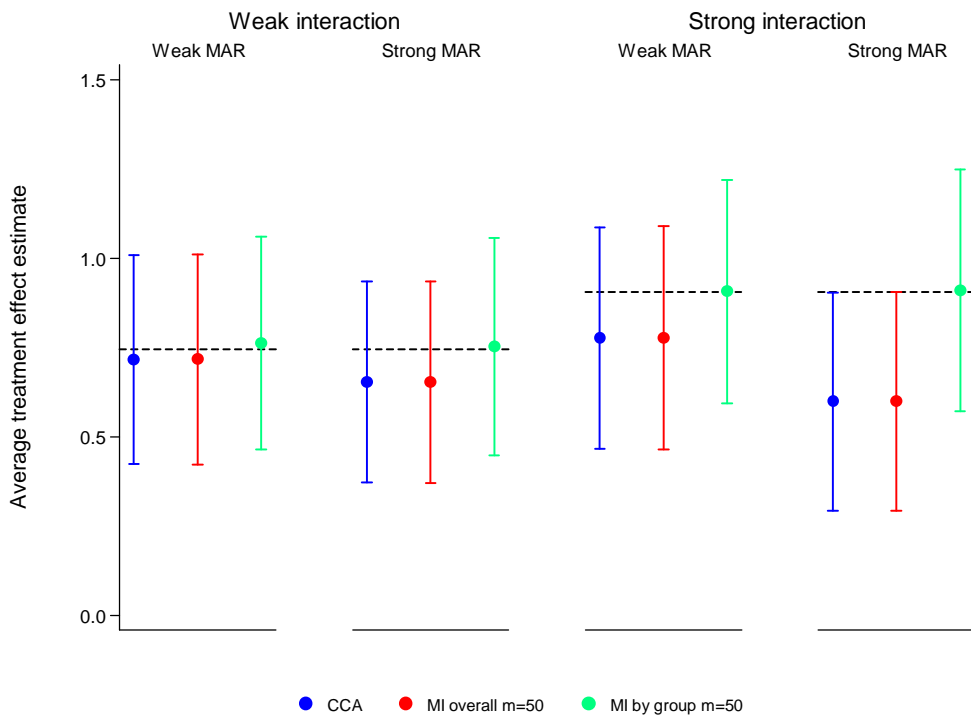


Figure 6.3. Mean average treatment effect estimates for 50% missing data in a binary outcome under the MAR X mechanism (odds of Y missing 1.5 (weak MAR) or 2.5 (strong MAR) times higher per standard deviation increase in X), incorrectly specified analysis model. Horizontal reference lines illustrate the least false average treatment effect in the absence of missing data. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

6.3.6. Missing data in a multivariate outcome

We now consider missing data in an outcome measured at repeated intervals following randomisation, where interest concerns the effect of treatment at the final time point. Unlike the univariate case, the validity of CCA is questionable in this setting since it cannot incorporate information from intermediate measures of the outcome. Such measures may be associated with the probability of missing data and the value of the outcome at the final time point. By exploiting information in partially observed cases, MI and likelihood-based approaches have been favored over CCA for the analysis of multivariate outcomes (13, 19, 47,

117). In what follows, we briefly introduce likelihood approaches for multivariate outcome data, describe the link between intermediate outcome measures and auxiliary variables, and present results from a simulation study comparing MI with alternatives.

Likelihood-based estimation of a linear mixed model (LMM) (79) is a popular alternative to MI for handling missing data in a multivariate outcome. Based on the multivariate normal distribution, this approach incorporates all observed information on the repeated measures of the outcome to produce estimates that are valid under a MAR assumption. No explicit imputation is involved. For outcomes collected at a limited number of fixed time points following randomisation, a LMM would typically include fixed effects for time (categorical), randomised group, and the interaction between randomised group and time. Within-subject dependence due to repeated measurements is accounted for through specification of a covariance structure. Several authors have recommended the unstructured covariance matrix since it is easily pre-specified, entails minimal power loss compared with more parsimonious choices (19, 118, 119) and ensures that estimates are approximately equivalent to and slightly more efficient than those obtained from a comparable MI procedure (11, 19). With a single intermediate measure Z , a LMM with adjustment for X is

$$\begin{pmatrix} Z_i \\ Y_i \end{pmatrix} \sim N \left\{ \begin{pmatrix} \alpha_0 + \alpha_1 T_i + \alpha_2 X_i \\ \beta_0 + \beta_1 T_i + \beta_2 X_i \end{pmatrix}, \begin{pmatrix} \sigma_Z^2 & \sigma_{ZY} \\ \sigma_{ZY} & \sigma_Y^2 \end{pmatrix} \right\}. \quad (2)$$

In applying MI, the repeated measurements of the outcome are usually treated as distinct variables in the imputation model. Where interest lies in the treatment effect at the final time point, the analysis model need not include the intermediate outcome measures; following imputation a comparison of final time point results is sufficient (120). In this case, the intermediate measures operate as auxiliary variables, assisting with the prediction of missing values at the final time point and making the MAR assumption more plausible. Other auxiliary variables, for instance measures of compliance or related outcomes, can also be added to the

imputation model as required. If data are collected but more likely to be missing following treatment discontinuation, an indicator variable for discontinuation may also be valuable as an auxiliary variable. The ability to incorporate auxiliary variables, both for univariate and multivariate outcomes, is considered one of the key strengths of MI (6). Less well known is that LMMs can also benefit from auxiliary variables through joint modelling with the outcome (19, 80). Using model (2) for illustration, Z could be an auxiliary variable rather than an intermediate outcome measure. By assuming an unstructured covariance matrix, multiple auxiliary variables are easily handled within a LMM (19).

For the simulation study, intermediate (Z) and final (Y) values of a continuous outcome were simulated from model (2) with $\beta_0 = \alpha_0 = 0$, $\beta_1 = \alpha_1 = 0.30$, and $\sigma_Z^2 = \sigma_Y^2 = 1$. To evaluate whether the correlation between Z and Y impacted on model performance, we considered $\sigma_{ZY} = 0.30$ or 0.70 . We also examined both weak (0.30) and strong (0.70) correlations between X and the outcome measures. Following generation of complete datasets, values in Y were set to missing such that the odds of missingness were λ times higher per SD increase in Z (with $\lambda = 1.5$ or 2.5 and for 20% or 50% missing data). In addition to CCA and MI, data were analysed using a LMM with an unstructured covariance matrix. Treatment effect estimates for LMMs in this article were obtained using restricted maximum likelihood estimation with degrees of freedom calculated according to the Kenward-Roger method (121).

MI overall, MI by group and the LMM produced unbiased treatment effect estimates across the 16 simulation scenarios (4 missing data scenarios \times 4 correlations), with coverage ≥ 0.94 throughout. Compared to the LMM, empirical SEs were on average 0.5% and 3.2% higher with MI overall and MI by group, respectively. The lost efficiency with MI by group was most noticeable in scenarios with 50% missing data and a strong MAR mechanism. Power was on average 0.3% lower for MI overall and 2.2% lower for MI by group compared to the LMM. By ignoring the intermediate measure of the outcome, CCA was, as expected, the least efficient approach. Although minimal in most settings, some

bias was also evident with CCA. Figure 6.4 illustrates performance in scenarios with 50% missing data, a strong MAR mechanism and where $\text{corr}(X, Y|T) = 0.30$. As seen in the figure, the relative performance of CCA was poor for $\sigma_{ZY} = 0.70$ (bias = 0.03, empirical SE 10.7% larger than the LMM). While outperforming CCA, MI offered no advantages over the LMM.

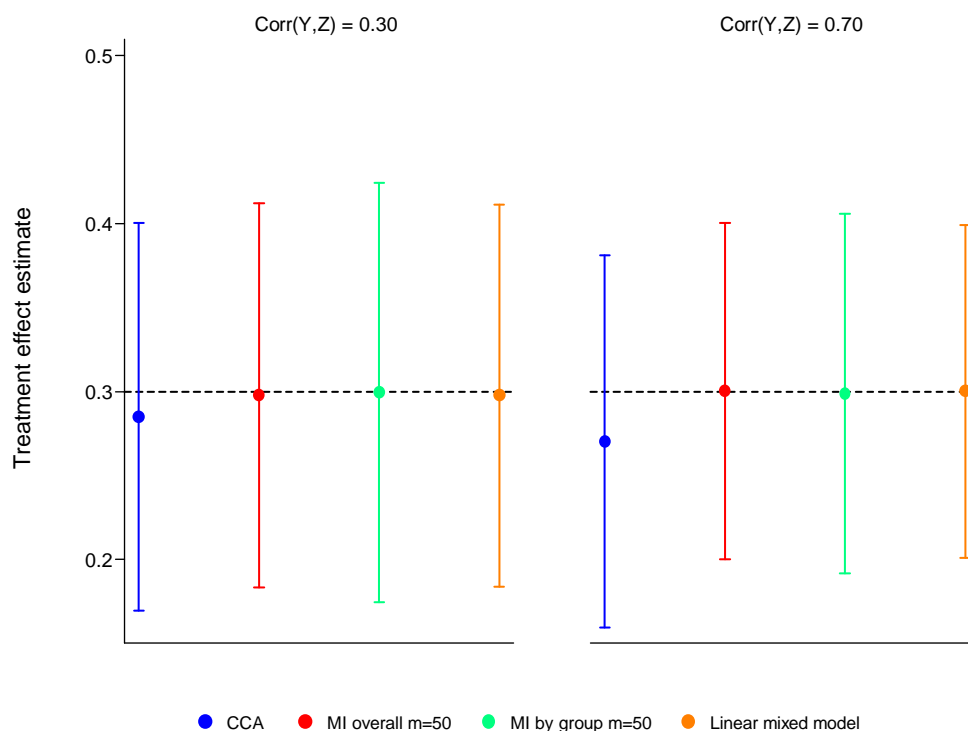


Figure 6.4. Mean treatment effect estimates for 50% missing data in a continuous multivariate outcome, $\text{corr}(X, Y|T) = 0.30$, strong MAR mechanism. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Similar results were obtained from a simulation study allowing missing data to occur in the intermediate as well as the final measure of the outcome, although the shortcomings of CCA were less pronounced in this setting (see web appendix B, Section 6.3.10; also available online at the journal website). We did not consider a simulation study for binary multivariate outcome data due to complexities in defining the estimand (see Section 6.3.9).

6.3.7. Missing data in a baseline covariate

Although missing baseline data can be avoided by requiring complete data collection before randomisation, this may not always be feasible (e.g. if a lengthy baseline interview is required). Unless baseline data are missing by design, it is implausible that missingness depends on randomised group given that baseline variables are measured before randomisation (19, 122). In this context, group comparisons based on complete cases should be unbiased, even if baseline data are MNAR. One potential limitation of the standard implementation of MI for imputing missing baseline data is that it ignores the independence of X and T . Chance imbalances in X in the observed data are incorrectly extrapolated to the missing data, which may result in a loss of efficiency (122). In this section, we evaluate the efficiency of MI using simulation, both for continuous and binary variables, and compare performance with alternative approaches.

Continuous baseline covariate and outcome

The binary indicator M_{Xi} for missing data in the baseline covariate X was first simulated with a probability of 0.20. Unlike other scenarios, we did not consider 50% missing data, since this degree of missingness seems unlikely for a baseline covariate pre-specified for adjustment. Next, baseline and outcome data were simulated from the models $X_i = \delta_X M_{Xi} + e_{1i}$ and $Y_i = 0.30T_i + \beta_2 X_i + \delta_Y M_{Xi} + e_{2i}$, with e_1 and $e_2 \sim N(0,1)$. The parameters δ_X and δ_Y in these models allow X and Y , respectively, to be associated with M_{Xi} . Both MCAR ($\delta_X = \delta_Y = 0$) and MNAR ($\delta_X = \delta_Y = 0.30$) mechanisms were considered in separate simulation scenarios. In choosing values for β_2 , we allowed $\text{corr}(X, Y | T, M_X = 0)$ to range between 0.10 and 0.90 in increments of 0.20.

In addition to MI and CCA, we evaluated the performance of mean imputation, the missing indicator method and a LMM with baseline as an outcome. In mean imputation, missing baseline values are replaced with the mean of the observed values across both groups (i.e. $X_i^* = \bar{X}_{obs}$ if $M_{Xi} = 1$). Although mean imputation

for addressing missing outcome data has been widely criticised for failing to incorporate missing data uncertainty (1, 71), overstated precision is not a concern in this setting given the independence of X and T and interest only in the effect of treatment (and not the effect of the covariate) (122). The missing indicator method involves mean imputation and the addition of a dummy variable indicating missing data to the analysis model (i.e. adding M_{Xi}). Despite being inappropriate for general use (123, 124), the missing indicator method has been validated for addressing missing covariate data in RCTs, where X and T are independent and missingness in X is conditionally independent of Y (122, 125). For strong correlations between X and Y , White and Thompson (122) found that mean imputation and the missing indicator method became more efficient when participants with missing data were given a weight of $1 - \text{corr}(X, Y|T, M_X = 0)^2$ in the analysis (with observed cases retaining a weight of 1). We investigated both unweighted and weighted approaches. For the LMM, we considered a joint model for X and Y , where X was assumed to be independent of T , i.e.

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N \left\{ \begin{pmatrix} \delta_0 \\ \beta_0 + \beta_1 T_i \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right\}.$$

Under both MCAR and MNAR mechanisms, all methods produced unbiased treatment effect estimates with nominal coverage throughout (range 0.94, 0.96). Despite this, noticeable differences in efficiency were apparent across the different approaches to handling missing data. Figure 6.5 summarises performance under the MCAR mechanism for $\text{corr}(X, Y|T, M_X = 0) = 0.10, 0.50,$ and 0.90 . As seen in the figure, CCA was close to optimal for a strong correlation between X and Y but inefficient for weak to moderate correlations. Both mean imputation and the missing indicator method performed well, with weighting becoming important for strong correlations. MI was marginally less efficient than the weighted approaches and the LMM (empirical SEs on average 0.3% larger), with little difference seen between MI overall and MI by group. Lastly unadjusted CCA was highly inefficient for moderate to strong correlations between X and Y . Efficiency results under the MNAR mechanism closely mirrored those of the

MCAR mechanism, with MI performing similarly to weighted mean imputation and the LMM across all values for $\text{corr}(X, Y|T, M_X = 0)$ (empirical SEs on average 0.3% larger with MI overall and MI by group than mean imputation and the LMM). Interestingly, the missing indicator method incorporating weights held a slight advantage over MI under the MNAR mechanism (empirical SEs on average 1.1% smaller than with MI), which can be attributed to inclusion of the prognostic variable M_X in the analysis model. A graphical summary of performance under the MNAR mechanism is shown in web appendix C (see Section 6.3.10; also available online at the journal website); we do not present results here given their similarity to the MCAR setting. Given the simplicity of alternative approaches to handling missing data in baseline covariates, there appears to be little reason to adopt MI in this setting.

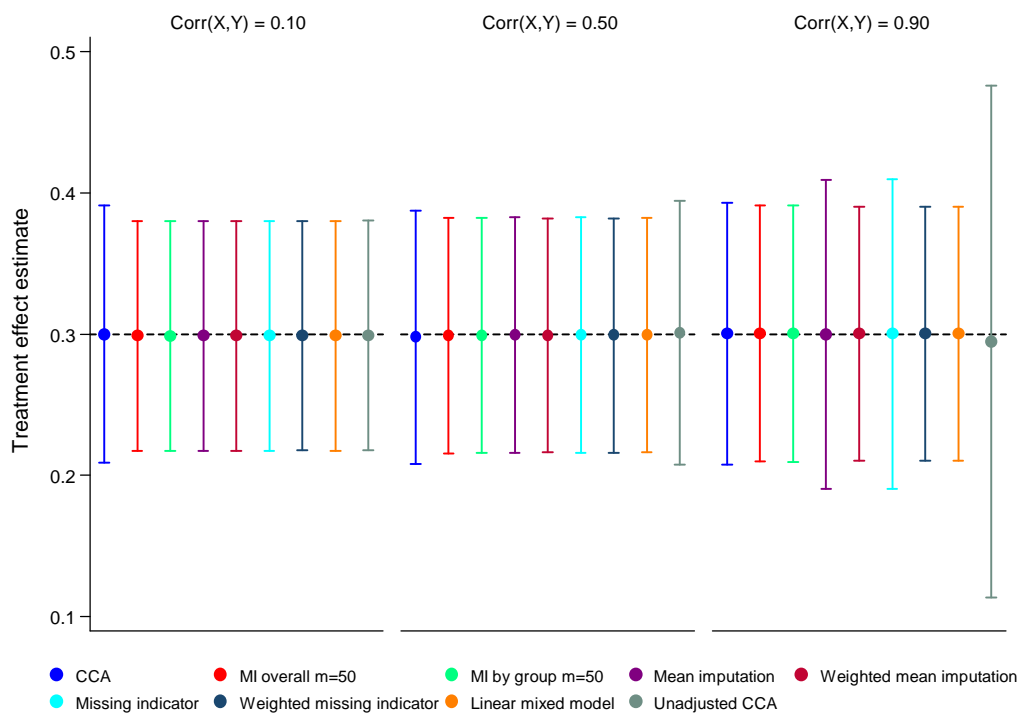


Figure 6.5. Mean treatment effect estimates for 20% missing data in a continuous baseline covariate, MCAR mechanism. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Binary baseline covariate and outcome

Following simulation of M_{Xi} with probability 0.20, baseline and outcome data were generated from the models $\text{logit } P(X_i = 1) = \delta_X M_{Xi}$ and $\text{logit } P(Y_i = 1) = \beta_0 + 0.69T_i + \beta_2 X_i + \delta_Y M_{Xi}$. The coefficient β_2 was varied so that $\text{OR}(Y, X|T, M_X = 0) = 2.0, 4.0, \text{ or } 8.0$, while β_0 was chosen to give $P(Y = 1|T = 0, M_X = 0) = 0.20$. Both MCAR ($\delta_X = \delta_Y = 0$) and MNAR ($\delta_X = \delta_Y = 0.69$) mechanisms were considered. We did not consider weighted methods or a LMM as in the continuous case, since these approaches are not applicable for binary outcomes.

Mean treatment effect estimates and empirical SEs for the MCAR mechanism are displayed in Figure 6.6. The clear outlier on these performance measures was unadjusted CCA. Since adjustment in logistic regression has the effect of increasing SEs and producing odds ratios that are further from the null (126), this finding is not surprising. Both MI overall and MI by group produced unbiased treatment effect estimates (absolute bias ≤ 0.004) with nominal coverage (range 0.95, 0.96) throughout, with little difference in empirical SEs between approaches. CCA produced treatment effect estimates with minimal bias, however empirical SEs were on average 10% larger than those of MI. For mean imputation and the missing indicator method, we observed a trade-off between efficiency and bias. For $\text{OR}(Y, X|T, M_X = 0) = 8.0$, both approaches exhibited modest efficiency advantages over MI (empirical SEs 4% smaller) at the expense of a small bias (-0.02) towards the null. In terms of average power, there were minimal differences between mean imputation (93.0%), the missing indicator method (93.0%) and the MI approaches (92.9%). The small bias of mean imputation and the missing indicator method arises because the methods estimate a treatment effect that lies between the marginal (unadjusted) and conditional (adjusted) estimands. As the proportion of missing data in X is increased, the methods shift from estimating the conditional estimand with no missing data to estimating the marginal estimand with no observed data (results not shown). Since for logistic regression the marginal estimand is always closer to the null, mean imputation and the missing

indicator method produce estimates of the conditional treatment effect that are biased towards the null.

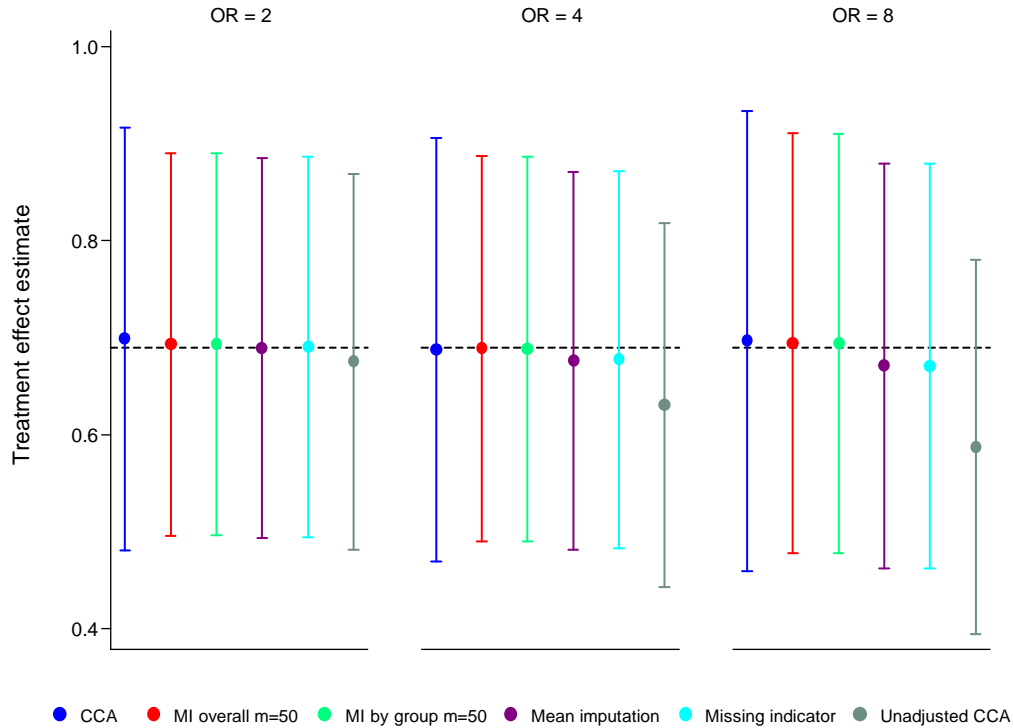


Figure 6.6. Mean treatment effect estimates for 20% missing data in a binary baseline covariate, MCAR mechanism. OR (odds ratio) refers to $OR(X, Y|T)$. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Although for $\delta_Y \neq 0$ the omission of M_X from analysis models changes the treatment effect estimated by logistic regression, the observed changes were minimal across the MNAR scenarios considered. Based on complete data, the “least false” treatment effect from a misspecified model omitting M_X was approximately 0.68 for all values of $OR(Y, X|T, M_X = 0)$. That distinction aside, results from the MNAR setting closely followed those of the MCAR setting (see Figure 6.7). In comparing MI with mean imputation, we once again observed a trade-off between efficiency and bias. For $OR(Y, X|T, M_X = 0) = 8.0$, the empirical SE of mean imputation was 4.7% smaller than MI, however the bias

was slightly more pronounced (-0.05 vs. -0.02). The missing indicator method performed similarly to mean imputation in terms of efficiency, however biases were smaller in magnitude with the missing indicator method due to correct specification of the analysis model. Excluding unadjusted CCA, all methods produced treatment effect estimates with correct coverage (range 0.94, 0.95).

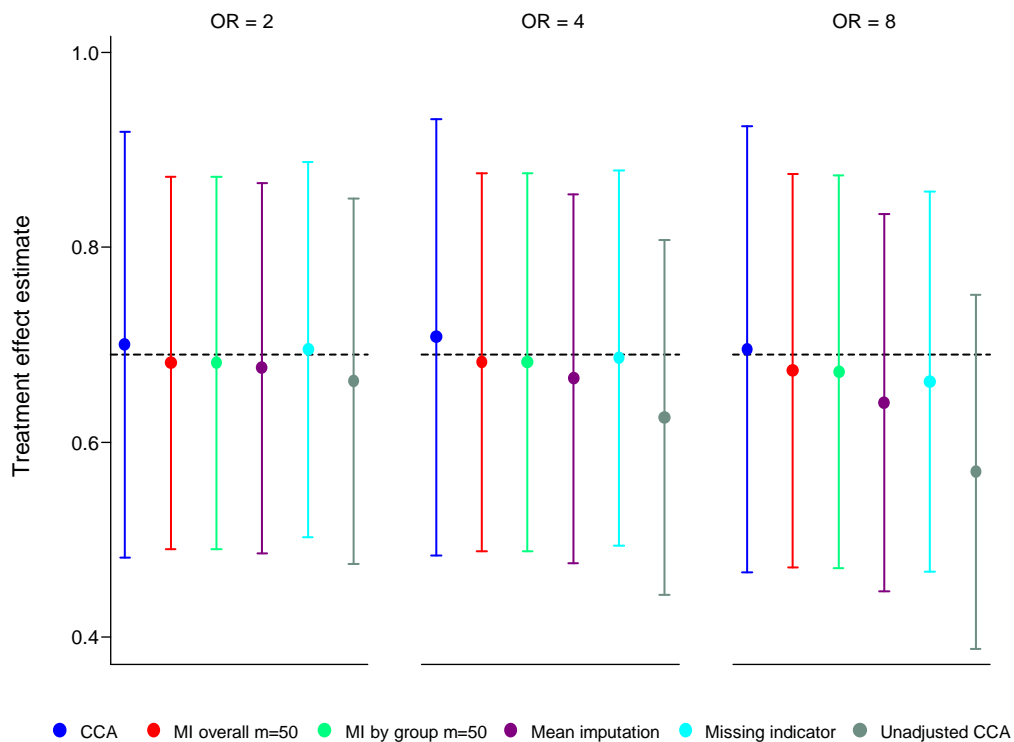


Figure 6.7. Mean treatment effect estimates for 20% missing data in a binary baseline covariate, MNAR mechanism. OR (odds ratio) refers to $OR(X, Y|T)$. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

6.3.8. Case study

The Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcome in Preterm Infants (DINO) trial was a blinded RCT conducted in five Australian hospitals between 2001 and 2007 (Australian New Zealand Clinical Trials Registry: ACTRN12606000327583). Preterm infants born < 33 weeks

gestation (n=657) were randomised to receive a high docosahexaenoic acid (DHA) or a standard DHA diet from within 5 days of commencing enteral feeds through to term. Randomisation was stratified by hospital, sex, and birth weight (<1250g, ≥1250g), with infants from a multiple birth randomised according to the sex and birth weight of the first born infant. Results for primary and key secondary outcomes have been published previously (91, 127, 128). In the primary trial publication (91), outcomes were re-analysed using MI following feedback from reviewers that all randomised infants had to be included in ITT analyses and that MI would be an appropriate approach to achieve this. To simplify the dataset for illustration purposes, second and subsequent born infants from a multiple birth and infants that died before term were ignored, resulting in an example dataset with 262 and 260 infants in the high and standard DHA groups, respectively.

To illustrate approaches for handling missing outcome data, we consider comparisons of fat free mass (FFM) at 7 years corrected age. Excluding two children that died after term, FFM was missing for 65/262 (24.8%) and 46/258 (17.8%) children in the high and standard DHA groups, respectively. Logistic regression analysis revealed differences between the five study centres in the odds of missing outcome data (global p-value = 0.03). No other predictors of missing data were identified. For predictors of the outcome, linear regression analysis revealed associations between FFM and centre, sex, and weight, height and systolic blood pressure at 7 years corrected age. Since centre and sex were baseline measures, for illustration purposes we imagine these variables were pre-specified as covariates for adjustment. Weight, height, and systolic blood pressure at 7 years corrected age were treated as auxiliary variables.

We estimated the effect of treatment using CCA, MI overall, MI by group, and a LMM. An unadjusted CCA was also conducted for comparison. Since the auxiliary variables contained missing data (approximately 10% for each variable), values were imputed using a Markov chain Monte Carlo algorithm assuming multivariate normality (11). Following a burn-in of 5000 iterations, $m = 50$

completed datasets were created. For the LMM, the three auxiliary variables and FFM were jointly modelled assuming an unstructured covariance matrix, with adjustment for centre and sex.

Treatment effect estimates are presented in Table 6.1. Although there was little evidence for an effect of treatment on FFM, subtle differences between the approaches are apparent. As expected, adjustment for prognostic baseline covariates in a CCA reduced the SE of the treatment effect estimate compared with the unadjusted analysis. By incorporating information from auxiliary variables, additional efficiency gains were evident for MI and the LMM, with similar estimates from the two approaches (as expected). However gains were small, perhaps because 48% of the children with a missing FFM value also had missing data on the three auxiliary variables. Even when fully observed, auxiliary variables may only have a meaningful impact on estimation when strongly correlated with the outcome (6, 48, 49).

Table 6.1. Treatment effect estimates for fat free mass (kg) at 7 years corrected age from the Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcome in Preterm Infants trial.

Method of analysis	Mean difference	Standard error	95% confidence interval
Unadjusted CCA	-0.007	0.259	-0.514 to 0.500
CCA	0.048	0.238	-0.420 to 0.515
MI overall m=50	-0.104	0.233	-0.562 to 0.353
MI by group m=50	-0.118	0.227	-0.563 to 0.327
Linear mixed model	-0.097	0.231	-0.551 to 0.356

Abbreviations: CCA, complete case analysis; MI, multiple imputation.

For missing data in a baseline covariate, we consider group comparisons of head circumference (HC) at term adjusted for birth HC. To focus on the problem of missing baseline data, 20 infants with missing outcome data were excluded from the analysis. Seven of these infants were missing birth HC and hence contained no information for estimating treatment effects, while the remaining 13 were assumed to be MAR and hence could be validly excluded (as demonstrated in Section 6.3.5). Of the remaining infants, birth HC was missing for 39/251 (15.5%)

and 42/251 (16.7%) in the high and standard DHA groups, respectively. Treatment effects were estimated using the same methods as in Section 6.3.7 for a continuous baseline covariate and outcome, with 50 imputations used for MI. In relation to the calculation of weights for mean imputation and the missing indicator method, in complete cases, the correlation between birth HC and HC at term was 0.43.

As illustrated in Table 6.2, estimates were similar across the nine statistical approaches. In line with simulation results for a moderate correlation between the baseline and outcome measure, CCA and unadjusted CCA produced the largest SEs for the effect of treatment. While outperforming CCA, MI did not offer any efficiency improvements over the remaining approaches.

Table 6.2. Treatment effect estimates for head circumference (cm) at term from the Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcome in Preterm Infants trial.

Method of analysis	Mean difference	Standard error	95% confidence interval
Unadjusted CCA	-0.060	0.136	-0.326 to 0.206
CCA	-0.058	0.134	-0.320 to 0.204
MI overall m=50	-0.023	0.125	-0.267 to 0.221
MI by group m=50	-0.027	0.125	-0.273 to 0.218
Mean imputation	-0.024	0.125	-0.269 to 0.221
Mean imputation with weights	-0.029	0.125	-0.274 to 0.215
Missing indicator	-0.028	0.125	-0.272 to 0.217
Missing indicator with weights	-0.032	0.124	-0.276 to 0.211
Linear mixed model	-0.029	0.125	-0.275 to 0.217

Abbreviations: CCA, complete case analysis; MI, multiple imputation.

Since the probability of missing baseline data differed across the five study centres, we considered additional sensitivity analyses where centre was added as a covariate in adjusted models and mean imputation was performed separately by centre. Although this resulted in small increases in precision compared to models ignoring centre, again MI did not outperform simpler approaches such as mean imputation and the missing indicator method with or without weights (SE = 0.123 for all approaches).

6.3.9. Discussion

In this article, we evaluated the performance of MI in the RCT setting. In line with theoretical results, in its standard implementation, MI produced unbiased treatment effect estimates when data were MAR and the analysis model was correctly specified. However, due to Monte Carlo simulation error, MI was often less efficient than alternative unbiased approaches. For missing outcome data, MI was less efficient than CCA for univariate outcomes and the LMM for multivariate outcomes. For missing data in a baseline covariate, MI failed to outperform methods such as mean imputation and the missing indicator method. As well as being less efficient, MI was generally more difficult to implement and took longer to run compared with alternatives. Being a stochastic analysis, it also had the disadvantage of not producing a unique treatment effect estimate. Given these limitations, we believe that MI should not be seen as the only acceptable way to address missing data in RCTs.

Collectively, our results underline the importance of context in choosing an approach for handling missing data. While MI is an extremely useful general purpose tool, it appears most beneficial in observational settings when there are missing data in confounding variables (93). In RCTs some of the value of MI is lost, and other approaches that are not widely recommended can be employed. For example, our simulation results confirm that mean imputation and the missing indicator method, whose use is ill-advised in most settings (1, 71, 123, 124), can be validly applied for addressing missing covariate data in RCTs. Similarly, despite general recommendations against the use of CCA (1, 71), it is optimal when missing data are restricted to a univariate outcome and variables associated with missingness are included as covariates in the analysis model (18, 23, 24). This scenario seems most pertinent to RCTs, where missing data tend to occur in the outcome. Of course should post-randomisation auxiliary variables for a univariate outcome be available, as is often the case, we then move into the setting of multivariate data and approaches such as MI or a LMM should be preferred over CCA.

Regarding choice of imputation strategy, we found that MI by group was slightly less efficient than MI overall for a correctly specified analysis model. However, when the analysis model overlooked an interaction effect involving randomised group, only MI by group produced unbiased estimates of the ATE. Thus in settings where MI is adopted, we recommend imputing by randomised group; compared to MI overall, this approach offers greater robustness at little cost. The approach is also consistent with general recommendations for over- rather than under-specifying imputation models (6, 11). It should be noted that imputing by group only protects against bias in estimating the ATE if effect modifiers are included in the imputation model. Another possibility is to include interaction terms in a single imputation model, but this approach is more complex and may not be obvious when analysis models do not include interaction terms. Although not considered in this article, we agree with previous recommendations for performing imputation separately by randomised group in settings involving subgroup analyses (13-15, 102).

Despite highlighting alternatives to MI in this article, we are not suggesting that it is inappropriate to use MI. To the contrary, we view MI as an attractive option given its considerable flexibility. It is not uncommon in RCTs for researchers to collect data on a large number of secondary outcomes. One of the strengths of MI is its ability to easily incorporate variables of different types (e.g. continuous, binary) in the imputation model, whether for univariate or multivariate data. An added benefit of including all outcomes in a single imputation model is that associations between related outcomes can aid imputation. Another appealing feature of MI is its ability to be implemented under an assumption that data are MNAR. This property makes MI well suited to undertaking sensitivity analyses around a primary assumption that data are MAR (76), and as a primary method of analysis in settings where data are believed to be MNAR. One such setting is RCTs where participants cannot followed up after discontinuing treatment. If all observed data are “on-treatment”, a MAR assumption entails estimating the effect of treatment had all participants remained on their assigned treatment (68). However, for a de facto type estimand (such as ITT), it may be more appropriate

to assume that data are MNAR. In this situation, reference-based sensitivity analyses have been proposed, which at present require the use of MI (112).

A limitation of the current study is that conclusions were based on a restricted set of simulation scenarios. Although we only considered simple randomisation to two groups, we anticipate that findings would extend to RCTs involving three or more randomised groups, unequal allocation probabilities, and randomisation using stratified blocks or minimisation. We also expect that our results for normally distributed and binary outcome variables would apply to most other outcome types. Three exceptions worth noting are time to event outcomes, where missing outcome data can be addressed via censoring, composite (scale) outcomes derived from multiple items, and binary multivariate outcomes. For missing data in a composite outcome, MI at the item level is a particularly convenient approach when the individual items are partially observed. Although likelihood-based alternatives for composite outcomes are also available (129), they are more difficult to implement. For binary multivariate outcomes, complexities arise due to differences between population-averaged and subject-specific estimands (130). Generalised mixed models can be implemented in a similar manner to LMMs for continuous data if subject-specific estimates are of interest (19); however, these models can be challenging to fit given the variety of estimation procedures available and the computational difficulties that can arise with large numbers of repeated measurements (131). MI is more appealing for producing population-averaged estimates (19).

A further limitation is that we did not consider the performance of inverse probability weighting (IPW). This approach, which involves weighting complete cases by the inverse of the probability of being a complete case, requires only a correctly specified model for the probability of missing data to produce valid estimates under a MAR assumption. However, IPW tends to be less efficient than MI and can be difficult to implement for non-monotone missing data patterns (26). Of relevance to the settings considered in this article, IPW is capable of producing population-averaged estimates for multivariate binary outcome data

and unbiased estimates of an ATE from a misspecified analysis model (26). IPW can also be appropriate in settings where data are missing by design and hence where the probability of being a complete case is known. We also did not evaluate multiple imputation, then deletion, which is a modification to standard MI where participants with imputed outcomes (but not imputed covariate values) are deleted from analysis datasets (8). The rationale behind this approach is that following imputation, participants with missing outcomes only contribute noise to the estimation procedure (8). Whether multiple imputation, then deletion is useful in the RCT setting is debatable however, since it is only applicable in settings where both covariate and outcome data are missing. Further, the approach should be avoided when auxiliary variables for the outcome are included in the imputation model (104), as is often the case.

In summary, MI is not the only option for handling missing data in RCTs. Although MI is appropriate in all contexts, simpler alternatives are often slightly superior. For missing outcome data, MI can be inferior to CCA and likelihood-based approaches, adding in unnecessary simulation error. For missing data in a baseline covariate, simpler approaches such as mean imputation and the missing indicator method can outperform MI. Should MI be adopted, we recommend imputing separately by randomised group.

6.3.10. Web appendix

Web appendix A. Missing data in a univariate binary outcome, where there is a correctly specified analysis model.

Considering $X_i = 0$ or 1 with probability 0.5 , binary outcomes were generated from the model $\text{logit } P(Y_i = 1) = \beta_0 + 0.69T_i + \beta_2X_i$, where the treatment effect of 0.69 corresponds to an odds ratio (OR) of 2.0 . To explore the impact that the strength of association between X and Y had on model performance, the coefficient β_2 was varied so that $\text{OR}(Y, X|T) = 2.0$ or 4.0 . Lastly the coefficient β_0 was chosen so that $P(Y = 1|T = 0) = 0.20$. Following the generation of

complete datasets, values in Y were set to missing according to the MAR X , MAR $X+T$ and MAR $X \times T$ mechanisms outlined in Section 6.3.5 (with X now a binary variable). Once again, λ (increase in odds of missing data per standard deviation increase in X) was set to 1.5 or 2.5 to indicate weak and strong missing data mechanisms, respectively, and both 20% and 50% missing data were considered.

CCA, MI overall and MI by group performed well in estimating the treatment effect for a binary outcome with missing data. Each method produced a mean treatment effect estimate of 0.70 across the 24 simulation scenarios (range 0.68, 0.71), with the small bias away from the null a product of the finite sample bias of logistic regression. CCA was the most efficient approach in all scenarios, with empirical standard errors on average 0.4% and 2.9% larger with MI overall and MI by group, respectively; differences were more pronounced with 50% missing data and under the strong MAR X and MAR $X+T$ mechanisms. Compared to CCA, power was on average 0.6% and 2.8% lower with MI overall and MI by group, respectively. Coverage for the three approaches remained close to 0.95 throughout (range 0.94, 0.96). Figure 6.8 illustrates performance in scenarios with 50% missing data, a strong MAR mechanism and for $OR(Y, X|T) = 4.0$. For reference, results are also displayed for unadjusted CCA. As seen in the figure, treatment effect estimates were noticeably different with unadjusted CCA, due both to inadequate handling of the missing data and the estimation of a different treatment effect in unadjusted logistic regression. For MI, empirical standard errors were marginally lower with MI overall than MI by group. As observed for continuous outcome data, MI offered no advantages over the simpler CCA in this setting.

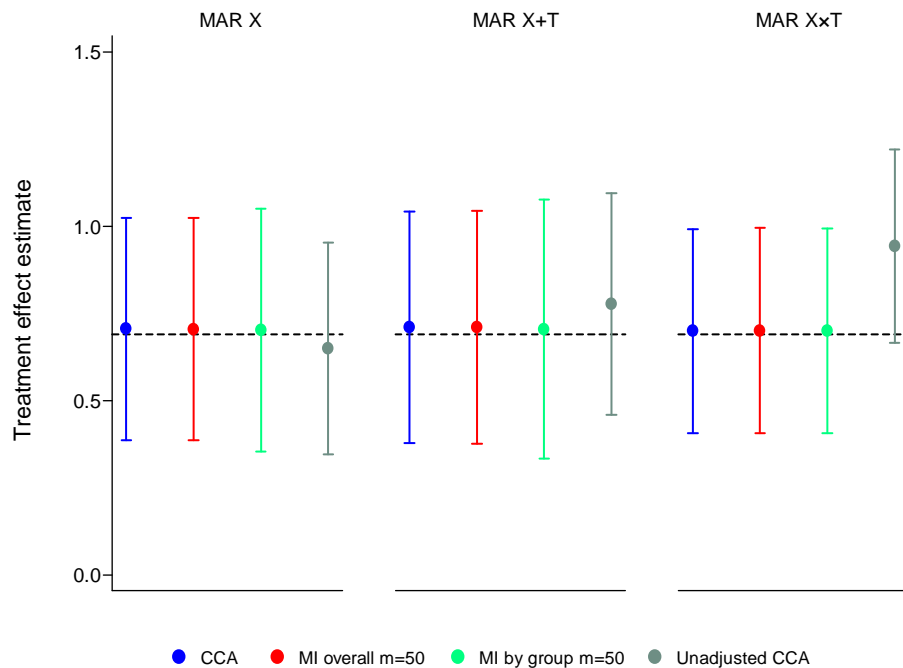


Figure 6.8. Mean treatment effect estimates for 50% missing data in a binary outcome, $OR(X, Y|T) = 4.0$, strong MAR mechanisms, correctly specified analysis model. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Web appendix B. Missing data in a continuous multivariate outcome where there is missing data in an intermediate measure of the outcome and the final outcome.

Intermediate (Z) and final (Y) values of a continuous outcome were simulated using the data generation model from Section 6.3.6 of the main article. Again we considered weak (0.30) and strong (0.70) values for σ_{ZY} and for the correlation between the baseline covariate X and the two outcome measures. Missingness was induced in a monotone pattern using two steps. In the first step, values in both Z and Y were simultaneously set to missing such that the odds of missingness were λ times higher in the control group and increased by a factor λ per SD increase in X (equivalent to the MAR X+T mechanism from Section 6.3.5). In a second step,

additional values in Y were set to missing such that the odds of missingness were λ times higher per SD increase in observed Z . We considered $\lambda = 1.5$ or 2.5 , with missing data proportions in (Z, Y) of $(0.10, 0.20)$ or $(0.25, 0.50)$. In addition to CCA and MI (using multivariate normal imputation), data were analysed using a LMM with an unstructured covariance matrix.

As expected for a MAR mechanism, MI overall, MI by group and the LMM produced unbiased treatment effect estimates with correct coverage (range 0.94, 0.46) across all scenarios. Compared to the LMM, empirical SEs were on average 0.4% and 3.0% higher with MI overall and MI by group, respectively (translating to average power losses of 0.1% and 2.8%). Although CCA was outperformed by MI and the LMM, deficiencies were not as pronounced as when missing data was restricted to the final outcome measure, as presented in Section 6.3.6. In fact, CCA was only marginally less efficient than the LMM for $\sigma_{ZY} = 0.30$ (empirical standard errors 0.9% larger). This is not an unexpected result, since Z has less information to contribute to estimation when it contains missing data. Figure 6.9 shows performance in scenarios with 50% missing data in Y , where $\text{corr}(X, Y|T) = 0.30$ and $\lambda = 2.5$. As seen in the figure, the shortcomings of CCA were most pronounced for $\sigma_{ZY} = 0.70$ (bias = 0.02, empirical SE 6.6% larger than the LMM), with little difference between MI and the LMM.

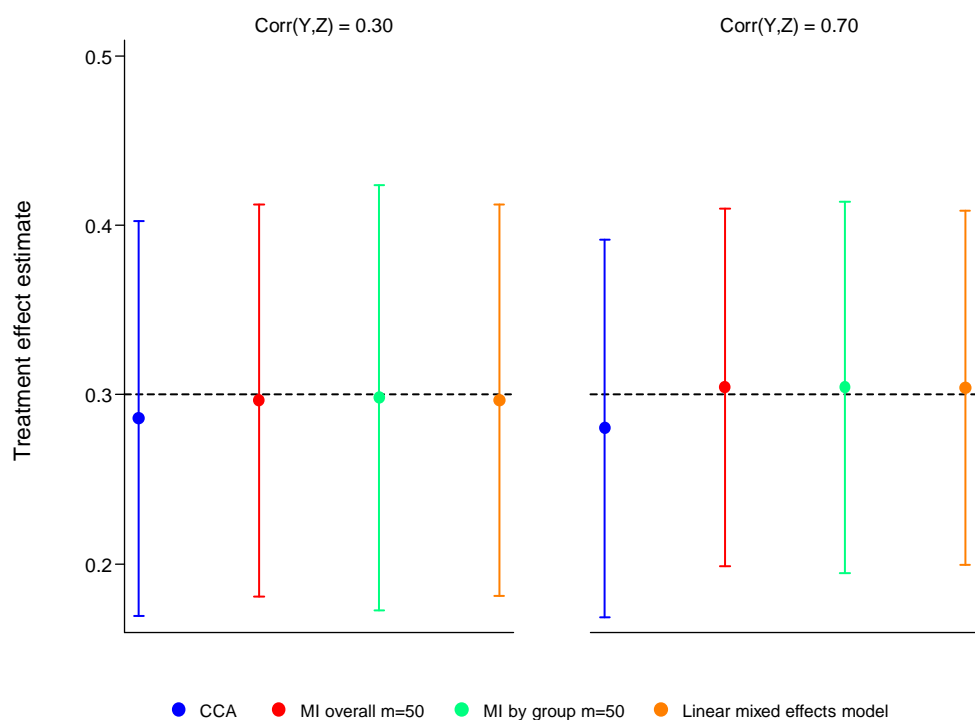


Figure 6.9. Mean treatment effect estimates for 25% and 50% missing data in an intermediate and final measure of a continuous multivariate outcome, respectively, $\text{corr}(X, Y|T) = 0.30$, strong MAR mechanism. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

Web appendix C. Missing data in a continuous baseline covariate, MNAR mechanism.

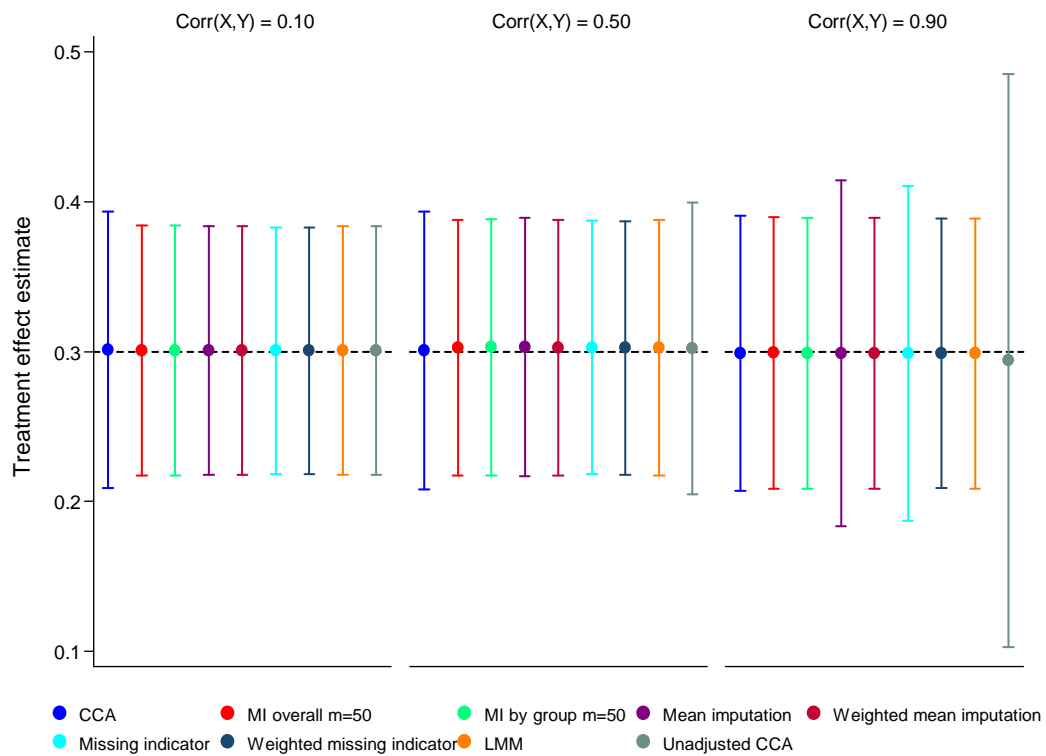


Figure 6.10. Mean treatment effect estimates for 20% missing data in a continuous covariate for adjustment, MNAR mechanism. Error bars correspond to empirical standard errors (± 1 standard error) across 2,000 simulated datasets.

*** End of published article ***

7. Multiple imputation in extended follow-up studies

7.1. Preface

This chapter contains the last of a series of four articles contributing to this thesis. The article, published in *Clinical Trials*, systematically reviews recently published extended follow-up studies of RCTs to summarise the extent and common sources of missing outcome data in this setting. The use of statistical approaches for handling missing outcome data in extended follow-up studies is also reviewed. Based on the findings of the systematic review, and using the DINO trial as a case study, some general recommendations for implementing MI in extended follow-up studies are provided at the conclusion of the chapter (see Section 7.4).

7.2. Statement of authorship

Title of paper	Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature.
Publication status	Published
Publication details	Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. <i>Clinical Trials</i> , 2017; 14(4): 387-95.

Principal author

Name (Candidate)	Thomas Sullivan		
Contribution	Designed the study, performed the literature search, extracted data, performed all analyses, interpreted the results, drafted the manuscript and acted as corresponding author.		
Overall percentage (%)	75		
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	11/07/2017

Co-author contributions

By signing the Statement of authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of co-author	Lisa Yelland		
Contribution	Contributed to the design of the study, extracted data, interpreted results and critically revised the manuscript.		
Signature		Date	11/07/2017

Name of co-author	Katherine Lee		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	26/07/2017

Name of co-author	Philip Ryan		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	12/07/2017

Name of co-author	Amy Salter		
Contribution	Contributed to the design of the study, interpreted results and critically revised the manuscript.		
Signature		Date	11/07/2017

7.3. Article

In this section, I provide the text, tables, figures, and appendices from the published manuscript.

7.3.1. Abstract

Background/Aims: After completion of a randomised controlled trial, an extended follow-up period may be initiated to learn about longer term impacts of the intervention. Since extended follow-up studies often involve additional eligibility restrictions and consent processes for participation, and a longer duration of follow-up entails a greater risk for participant attrition, missing data can be a considerable threat in this setting. As a potential source of bias, it is critical that missing data are appropriately handled in the statistical analysis, yet little is known about the treatment of missing data in extended follow-up studies. The aims of this review were to summarise the extent of missing data in extended follow-up studies and the use of statistical approaches to address this potentially serious problem.

Methods: We performed a systematic literature search in PubMed to identify extended follow-up studies published from January to June 2015. Studies were eligible for inclusion if the original randomised controlled trial results were also published and if the main objective of extended follow-up was to compare the original randomised groups. We recorded information on the extent of missing data and the approach used to treat missing data in the statistical analysis of the primary outcome of the extended follow-up study.

Results: Of the 81 studies included in the review, 36 (44%) reported additional eligibility restrictions and 24 (30%) consent processes for entry into extended follow-up. Data were collected at a median of 7 years after randomisation. Excluding 28 studies with a time to event primary outcome, 51/53 studies (96%) reported missing data on the primary outcome. The median percentage of

randomised participants with complete data on the primary outcome was just 66% in these studies. The most common statistical approach to address missing data was complete case analysis (51% of studies), while likelihood-based analyses were also well represented (25%). Sensitivity analyses around the missing data mechanism were rarely performed (25% of studies), and when they were, they often involved unrealistic assumptions about the mechanism.

Conclusions: Despite missing data being a serious problem in extended follow-up studies, statistical approaches to addressing missing data were often inadequate. We recommend researchers clearly specify all sources of missing data in follow-up studies and use statistical methods that are valid under a plausible assumption about the missing data mechanism. Sensitivity analyses should also be undertaken to assess the robustness of findings to assumptions about the missing data mechanism.

7.3.2. Introduction

After a randomised controlled trial (RCT) has come to its protocol defined end, it may be desirable to instigate an extended follow-up period to learn about longer term impacts of the intervention. In prevention and treatment trials, extended follow-up can be important for verifying that early effects on biomarkers of disease activity translate to longer term effects on more clinically meaningful endpoints (83, 84). In perinatal trials, extended follow-up may be initiated to evaluate impacts on development in later childhood. In other settings, investigators may choose to follow up participants to learn more about disease progression, long-term safety, treatment-related costs, or the maintenance of early effects (16). The key benefit of initiating extended follow-up is the cost saving associated with using an already established cohort. Given the substantial investment required in designing a trial, recruiting participants, providing treatment and collecting baseline data, it is not surprising that many trials do eventually transition to extended follow-up studies (16).

Like standard RCTs, missing data can threaten the validity of findings from extended follow-up studies. The process of transitioning to an extended follow-up study may occur years after completion of the original RCT, leaving the task of re-contacting participants and obtaining outcome data particularly difficult. Even without a delay in commencing extended follow-up, the longer time between randomisation and final outcome assessment may be associated with greater participant attrition. In multicentre trials, some centres might not participate in extended follow-up, or investigators could impose other eligibility restrictions for inclusion into extended follow-up. Depending on the information provided to participants in the original RCT, a separate consent form for extended follow-up may also be necessary. Some participants may be unwilling to consent at this stage. Finally, participants may simply fail to provide information about a particular measure during extended follow-up. Taken together these varied sources of missing data (attrition over time, ineligibility, non-consent and item non-response) could result in a large proportion of the original randomised cohort having missing outcome data.

The most effective way to minimise the impact of missing data in extended follow-up studies is to prevent it. In a recent review, Drye et al. (16) considered logistical issues in undertaking extended follow-up studies, with several of their recommendations focusing on the prevention of missing data. Suggestions included minimising the time between trial completion and follow-up study commencement, maintaining participant contact details at a central facility, informing participants about possible future contact, and attempting to contact participants who were unable to complete the original RCT. Even with the most rigorous planning, however, there will invariably be some missing data in extended follow-up studies. Since inadequate treatment of missing data in an analysis can result in substantial bias and inefficiency (1), it is critical that appropriate statistical methods are adopted.

The validity of any statistical method used to handle missing data depends on the mechanism responsible for the data being missing. Using Rubin's system (2), data

can be classified as missing completely at random if missingness is independent of observed and unobserved data, missing at random if missingness is independent of unobserved data given observed data, and missing not at random if missingness is dependent on unobserved data given observed data. Since the mechanism cannot be verified from observed data, researchers are encouraged to state and justify the assumption made about the missing data mechanism in the main analysis and to undertake sensitivity analyses around this assumption (1, 13, 17, 20).

A common approach to handling missing data in RCTs is to perform a complete case analysis (109, 132), which involves restricting the analysis to participants with complete data on all variables in the analysis model. Although simple to implement, complete case analysis is often inefficient and can introduce bias when data are not missing completely at random (1, 14). Single imputation methods, which involve replacing missing values with single imputed values, are also commonly used in RCTs (109, 132). A major concern with the application of these methods is that analyses are often incorrectly conducted as if all data were observed, which can lead to overstated precision (1, 13). A noteworthy single imputation method for longitudinal settings is the last observation carried forward, where missing outcomes are replaced by the last observed measurement. As well as concerns around overstated precision, this method can introduce bias when outcome values change following the last observed measurement (13).

Several more principled alternatives for handling missing data are available. Inverse probability weighting, where complete cases are weighted by the inverse of the probability of being a complete case (26), and likelihood-based methods (e.g. mixed models for repeated measures data) (79, 117) produce valid inference under a missing at random assumption. Another approach typically implemented under a missing at random assumption is multiple imputation (4), although application under missing not at random mechanisms is also possible. In its standard implementation, multiple imputation involves replacing each missing observation with multiple independent draws from the posterior predictive

distribution of the missing data conditional on the observed data, a process that generates multiple complete datasets. Following analysis, results for each complete dataset can be combined using appropriate rules to give a single estimate. An alternative to multiple imputation is model-based single imputation; however, special methods such as the jackknife are required to obtain valid standard error estimates (133).

Given recommendations for the use of inverse probability weighting, likelihood-based methods, and multiple imputation in guidance documents for RCTs (1, 71), it seems reasonable that these methods should be preferred in extended follow-up studies. Yet implementation in this setting may be more complex given the additional sources of missing data present. Consider an extended follow-up study involving a separate consent process and where eligibility is restricted to participants who completed the original RCT, as illustrated in Figure 7.1. The analysis could include all randomised, all eligible, or all consenting participants. Incorporating the full randomised cohort in the analysis preserves the benefits of randomisation, but there may be a large amount of missing data to account for and a mixture of missing data mechanisms at play, since reasons for missing data could differ between ineligible participants, non-consenters, and consenters. Satisfying an assumption about the missing data mechanism might be more feasible if calculations only incorporate eligible or consenting participants, but then the benefits of randomisation are diminished. The population of interest for the chosen measure of intervention effect (e.g. all randomised participants for intention to treat) should also be taken into account when choosing a participant group to incorporate in the analysis. Discussion of these issues is lacking in the literature, and it is unclear how missing data in this context are being handled in practice.

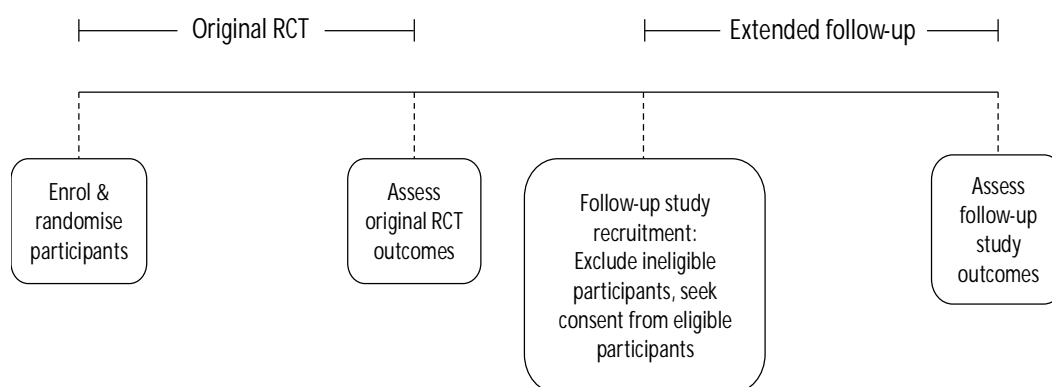


Figure 7.1. Timeline of an extended follow-up study.

We undertook a systematic review of the literature to investigate the treatment of missing data in extended follow-up studies. Although reviews on the treatment of missing data have been undertaken for other research designs, for example randomised trials (12, 109, 132) and cluster randomised trials (134, 135), to our knowledge, this is the first review of missing data in extended follow-up studies. The aims of the review were to summarise the extent of missing data in follow-up studies, the quality of reporting around missing data, and the use of statistical approaches to address this potentially serious problem.

7.3.3. Methods

Research articles published in English between 1 January 2015 and 30 June 2015 were reviewed. Studies were eligible for inclusion if the original trial results were also published and if the main objective of extended follow-up was to compare the original randomised groups. Since options for handling missing data are limited in small sample sizes, only studies involving the randomisation of at least 100 participants were considered eligible. Articles were excluded if the original trial was a pilot or dose-finding study, if extended follow-up was pre-specified in the original trial protocol, or if the article included multiple trial cohorts (as this would lead to additional complexities in handling missing data). Where multiple articles reporting on the same follow-up study were discovered, only the first published article was included in the review to avoid replicating results.

The search was conducted in PubMed on 18 January 2016. Search terms were based on the Cochrane sensitivity and precision maximising search strategy for identifying randomised trials (136), with additional terms for “follow-up”, “continuation study” and “long-term effects”. The search strategy is detailed in Table 7.1. Titles and abstracts of identified articles were examined and classified as potentially eligible or ineligible. Full texts of potentially eligible articles were then examined to confirm eligibility, with information from eligible articles transcribed to a pre-piloted data extraction form developed specifically for this review. Details reported in Supplementary Materials and Web Appendices were included in this review process. The assessment was carried out in full by one reviewer (T.R.S.), with a second reviewer (L.N.Y.) independently examining 20% of the articles. Interrater agreement for article eligibility, as indicated by a Kappa statistic (137), was estimated to be 0.89. All disagreements were resolved by discussion.

Table 7.1. Search strategy to identify extended follow-up studies (PubMed search date 18 January 2016).

(randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized[tiab] OR randomised[tiab] OR placebo[tiab] OR clinical trials as topic [mesh: noexp] OR randomly[tiab] OR trial[ti]) NOT (animals[mh] NOT humans[mh]) AND ("follow-up" [ti] OR "followup" [ti] OR "continuation study" [ti] OR "long-term effects" [ti]) AND ("2015/01/01"[PDat] : "2015/06/30"[PDat]) AND English[lang]

For each eligible follow-up study, key details about the original RCT were recorded, including the unit of randomisation (individuals vs. clusters), number of randomised participants, number of treatment arms, and type of intervention. Information on the use of separate eligibility restrictions and consent processes for entry into extended follow-up was then documented, including numbers eligible and consenting where applicable. The duration of time between randomisation or completion of the original RCT and completion of the extended follow-up study was also recorded.

In extracting information about the handling of missing data, the review focused on the analysis of a single primary outcome. If multiple primary outcomes were identified in the article, the primary outcome of interest was defined as the first primary outcome used to determine the sample size, or the first primary outcome identified otherwise. If no primary outcome was identified, it was taken to be the first outcome used to justify the sample size, otherwise the first outcome presented in a table or figure. With the exception of time to event outcomes, where missing data can, in part, be addressed through censoring, the number of complete cases for the primary outcome was recorded. For outcomes measured repeatedly over time, the number of complete cases was taken to be the number available at the final assessment. In studies with missing data on the primary outcome, the following information was extracted: measure of intervention effect of interest (e.g. intention to treat), statement and justification of the missing data mechanism assumed, and statistical method used to handle missing data in both the main analysis and in sensitivity analyses (if performed).

7.3.4. Results

The electronic search identified 420 articles, of which 274 were excluded based on a review of titles and abstracts. Of the remaining 146 articles, 81 satisfied eligibility criteria and were included in the review (Figure 7.2). The full list of included articles is provided in the web appendix (see Section 7.3.6; also available online at the journal website).

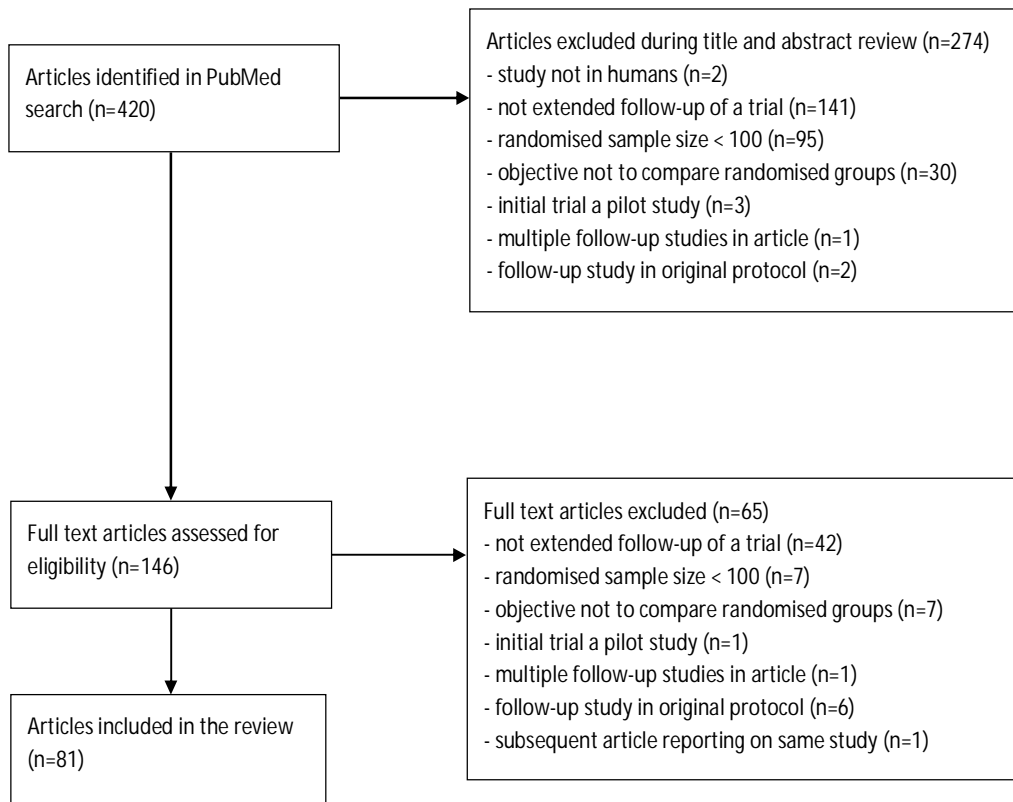


Figure 7.2. Flow diagram for systematic review.

Characteristics of original trial

Key characteristics of the original trial are presented in Table 7.2. The median number of randomised participants was 299, with the majority of trials having two treatment arms and randomising individuals rather than clusters. The most common type of intervention was a drug or medical device (41%), with surgical (12%), psychological (11%) and nutritional supplement (10%) interventions also well represented. Reporting of blinding was poor, with 56% of articles providing insufficient detail to determine the type of blinding employed. In many articles, readers were referred to the original RCT publication for details on trial design.

Table 7.2. Characteristics of the original trials^a.

Characteristic	Number of studies (n=81)
Number of participants: median (inter)	299 (154, 1152) [n=80]
Number of treatment arms	
2	65 (80%)
3 or more	16 (20%)
Randomisation	
Individual	67 (83%)
Cluster	8 (10%)
Unclear	6 (7%)
Intervention	
Drug/device	33 (41%)
Exercise/diet/lifestyle	7 (9%)
Nutritional supplement	8 (10%)
Psychological	9 (11%)
Surgical	10 (12%)
Other	14 (17%)
Blinding	
Unblinded	14 (17%)
Participants blinded	1 (1%)
Outcome assessors blinded	14 (17%)
Both participants and outcome assessors blinded	7 (9%)
Unclear	45 (56%)

Abbreviations: IQR, interquartile range.

^a Values are n(%) unless otherwise indicated.

Characteristics of extended follow-up study

Information on the transition to extended follow-up is presented in Table 7.3. Of the 81 articles included in the review, 36 (44%) reported separate eligibility restrictions for entry into extended follow-up. The most common restriction concerned satisfactory adherence to the protocol in the original trial (22 studies). Participants were also ruled ineligible according to their enrolling centre (three studies), treatment arm (two studies), or other baseline (three studies) or post-randomisation (six studies) characteristics. Across studies reporting eligibility restrictions, the median percentage of randomised participants eligible for follow-up was 86%. A total of 24 studies (30%) reported using a separate consent process for follow-up; the median percentage of randomised participants providing consent was 70% in these studies. It is possible that other studies used eligibility restrictions and consent processes for follow-up but failed to report them. Overall,

the median duration of time from randomisation to completion of extended follow-up was 84 months, representing a median of 52 months of additional follow-up beyond the original RCT.

Table 7.3. Characteristics of the extended follow-up studies^a.

Characteristic	Number of studies (n=81)
Reported on an eligibility restriction for the follow-up study	
Yes	36 (44%)
No	45 (56%)
Percentage of randomised participants eligible: median (IQR)	85.6 (73.1, 91.9) [n=32]
Reported on the use of a separate consent process for the follow-up study	
Yes	24 (30%)
No	57 (70%)
Percentage of randomised participants consenting: median (IQR)	70.3 (54.2, 77.2) [n=20]
Months from randomisation to follow-up study completion: median (IQR)	84 (38, 120) [n=77]
Months from original RCT completion to follow-up study completion: median (IQR)	52 (24, 72) [n=72]
Same primary outcome as in original RCT but at a later time point	
Yes	17 (21%)
No	22 (27%)
Unclear	42 (52%)
Type of primary outcome in follow-up study	
Continuous	36 (44%)
Binary	14 (17%)
Categorical	3 (4%)
Time to event	28 (35%)
Number of measurements on primary outcome ^b	
Single	17 (32%)
Repeated	36 (68%)
Accounted for missing data in sample size calculations ^c	
Yes	5 (10%)
No	1 (2%)
No calculation provided	45 (88%)
Reported information on the amount of missing data ^c	
By treatment arm	47 (92%)
Overall only	4 (8%)
Percentage complete cases among all randomised: median (IQR) ^c	65.9 (53.1, 78.6) [n=48]

Abbreviations: IQR, interquartile range; RCT, randomised controlled trial.

^a Values are n(%) unless otherwise indicated.

^b Excluding n=28 studies with a time to event primary outcome.

^c Excluding n=28 studies with a time to event primary outcome and n=2 studies that did not report missing data.

Table 7.3 also provides details on the primary outcome of extended follow-up. In 17 studies (44%), the primary outcome was unchanged from the original RCT, albeit measured at a later time point. The majority of studies involved either a

continuous (44%), time to event (35%), or binary (17%) primary outcome. Excluding 28 studies with a time to event primary outcome, 51/53 studies (96%) reported missing data on the primary outcome. In one study, all randomised participants had primary outcome data available, while another study provided insufficient details to determine if there were missing data. Of the 51 studies reporting missing data, only five accounted for missing data in sample size or power calculations; 45 did not provide any justification for the sample size in the extended follow-up study at all. Reporting on the extent of missing data was fairly rigorous, with 47/51 studies reporting some information on missing data proportions separately by treatment arm. Across treatment arms, the precise number of complete cases on the primary outcome was presented or possible to infer in 48/51 studies. Among these studies, the median percentage of randomised participants with complete data on the primary outcome was just 66% (interquartile range 53%, 79%).

Handling of missing data in the main analysis

Of the 51 studies reporting missing data on the primary outcome, 26 (51%) failed to identify the measure of intervention effect, or estimand, of interest (Table 7.4). A total of 18 studies undertook analyses according to the intention to treat principle and three according to a per-protocol approach. The remaining four studies defined some other estimand of interest. Of the 18 studies reportedly following the intention to treat principle, six restricted eligibility for extended follow-up according to adherence to the protocol in the original RCT. Only 5/51 studies with missing data explicitly stated the missing data mechanism assumed in the main analysis, with just three of these providing justification for the assumption. In two studies, a missing at random assumption was deemed plausible since baseline characteristics were similar between participants with missing and observed outcomes (suggesting data might have been missing completely at random). Another study identified predictors of missing data and justified a missing at random assumption by incorporating these predictors into a likelihood-based linear mixed model.

Table 7.4. Analysis of the primary outcome^a.

Characteristic	Number of studies (n=51) ^b
Measure of intervention effect (estimand) of interest	
Intention to treat	18 (35%)
Per protocol	3 (6%)
Other	4 (8%)
Not stated	26 (51%)
Reported on missing data mechanism assumed in the analysis	
Missing at random	5 (10%)
Not stated	46 (90%)
Justified the assumption made about the missing data mechanism	
Yes	3 (6%)
No	2 (4%)
Not applicable	46 (90%)
Statistical approach	
Complete case analysis	26 (51%)
Single imputation	3 (6%)
Multiple imputation	4 (8%)
Likelihood based	13 (25%)
Estimating equation method	4 (8%)
Unclear	1 (2%)
Performed a sensitivity analysis around the missing data mechanism	
Yes	13 (25%)
No	38 (75%)

^a Values are n(%).

^b Excluding n=28 studies with a time to event primary outcome and n=2 studies that did not report missing data.

The most common approach for handling missing data in the main analysis was to perform a complete case analysis (26/51 studies; Table 7.4). Of studies using complete case analysis, 17/26 (65%) involved a primary outcome that was measured repeatedly over time, and so analyses (final time point analysis or analysis of variance (ANOVA)) may have excluded participants with available data on earlier measurements. Among the nine studies using complete case analysis for a univariate (once-measured) primary outcome, only one adjusted for baseline covariates. A total of 13 studies used a likelihood-based approach in the main analysis, 10 involving a repeatedly measured outcome and 3 a univariate outcome with clustering in the data. Single imputation methods were used in three studies: two used last observation carried forward and one assumed that participants were disease free if data were missing. Four studies used estimating equations (without probability weights) to account for clustering, which relies on data being missing completely at random (138). In one study, the method of

analysis was unclear, while four studies used multiple imputation to address missing data. In three studies using multiple imputation, imputed datasets included all randomised participants. In the other, where interest concerned the intention to treat estimand, data were only imputed for participants who consented to extended follow-up, although imputation back to the full randomised cohort was explored in a sensitivity analysis. Regarding method of multiple imputation, two studies used chained equations to generate 100 imputed datasets, with additional predictors of the outcomes and of missing data (known as auxiliary variables) included in the imputation model. One study used an expectation maximisation algorithm to generate 20 imputed datasets, while another did not report any details of the imputation methodology.

Sensitivity analysis

Of the 51 follow-up studies with missing data, 13 (25%) reported undertaking sensitivity analyses where an alternative statistical method was used to address missing data. Methods of sensitivity analysis included complete case analysis (five studies), likelihood-based (three studies), multiple imputation (two studies), last observation carried forward (two studies), other single imputation procedure ignoring missing data uncertainty (two studies), and estimating the range of possible treatment effects for missing data in a binary outcome (one study). Of note, only six studies were judged to have made a less restrictive assumption about the missing data mechanism in sensitivity analyses, with just two considering missing not at random mechanisms. Eight studies presented treatment effect estimates along with confidence intervals or standard errors from sensitivity analyses, one graphically presented a range of p-values, while the remaining four only commented that conclusions were unchanged following sensitivity analyses.

7.3.5. Discussion

In this article, we reviewed the occurrence and treatment of missing data in recently published extended follow-up studies. With primary outcome data

collected at a median of 7 years after randomisation, and with many studies reporting separate eligibility restrictions and consent processes for entry into extended follow-up, it was not surprising to find high proportions of missing data. We found that the median percentage of randomised participants with complete data on the primary outcome was just 66%. In comparison, systematic reviews of RCTs have reported median complete data proportions in the vicinity of 90% (range 89 to 92%) (7, 12, 109). Despite the considerable threat of missing data, several weaknesses in the statistical handling of missing data were identified. Only half the included articles reported the estimand of interest, less than 10% explicitly stated the missing data mechanism assumed in the analysis and just 25% undertook sensitivity analyses around the missing data mechanism. Further, roughly 60% of studies performed the main analysis under the strong assumption that data were missing completely at random. Clearly there is room for improvement.

Complete case analysis was the most popular statistical approach in this review, despite criticism in guidance documents for RCTs (1, 71). It is possible that researchers remain unclear about the underlying assumptions required to guarantee the validity of this approach. Indeed, one study employing complete case analysis claimed they made “no assumptions about missing data” (139). For outcomes measured repeatedly over time, there is little justification for complete case analysis. Even in the unlikely scenario that repeated measures data are missing completely at random, complete case analysis tends to be inefficient since participants with intermediate measures on the outcome can be excluded from the analysis. For univariate outcomes, complete case analysis may be more defensible. Research has shown that complete case analysis with covariate adjustment produces unbiased and efficient estimates of regression parameters when univariate outcome data are missing at random conditional on covariates (18, 23, 24). In the context of extended follow-up studies, this means that inference can be improved by identifying and subsequently adjusting for baseline predictors of missing data. Unfortunately we only observed one study where

baseline variables were adjusted for in a complete case analysis of a univariate outcome.

Encouragingly, likelihood-based approaches or multiple imputation were used in 33% of the included studies, which is higher than usage rates of around 25% observed in recent systematic reviews of RCTs (109, 132). It was also promising to find only two studies that used last observation carried forward in the main analysis. This method relies on the questionable assumption that responses remain stable beyond the last observed measurement. Even in settings where this assumption is plausible, the approach tends to produce confidence intervals that are too narrow (1, 13, 110).

As well as choosing and justifying a sensible method of analysis in the presence of missing data, it is critical that researchers explicitly state the estimand of interest (1). Since discussion on the topic is still evolving (66, 68, 140), we avoid trying to define here exactly what constitutes an estimand, yet it remains clear that one must know what is being estimated to judge the appropriateness of a statistical approach. Unfortunately less than half of the included studies stated the estimand of interest. Among studies where it was stated, the majority focused on the intention to treat estimand. Interestingly, three studies undertaking intention to treat analyses used a complete case analysis for repeated measures data, and so may have excluded participants with available outcome data at earlier time points. In addition, six studies restricted eligibility for entry into extended follow-up according to adherence to the protocol in the original RCT. Both these approaches ignore recommendations for undertaking intention to treat analyses, which specifically advocate using all available outcome data in the analysis and attempting to collect outcome data on all randomised participants (1, 20, 68).

Another major shortcoming identified in the review concerned the use of sensitivity analyses around the missing data mechanism. Sensitivity analyses were infrequently performed, and when they were, they often involved strong assumptions about the missing data mechanism. Only two studies considered

missing not at random mechanisms in sensitivity analyses. Guidance documents for RCTs have emphasised the need to consider plausible alternative assumptions about the missing data in sensitivity analyses (1, 20, 71), typically by relaxing the assumption about the missing data mechanism. Given the high levels of missing data observed in this review, we believe these recommendations are especially relevant in extended follow-up studies.

Contrary to expectations, we did not find any discussion on the merits of incorporating the full randomised cohort in the analysis compared with a subsample in follow-up studies involving separate eligibility restrictions and/or consent processes. Although one study employing multiple imputation included consenting participants in the main analysis and the full randomised sample in a sensitivity analysis, the rationale for this approach was not described. In encouraging researchers to adopt principled approaches such as inverse probability weighting, likelihood-based methods and multiple imputation, some guidance around the choice of participant group to incorporate in such an analysis, and factors that might influence this decision, would be a welcome contribution.

A limitation of this review is that for feasibility we extracted information only from published follow-up studies and associated supplementary materials. Further details could have been obtained from the original RCT publication, in published protocols or by contacting authors. Other approaches to addressing missing data may have been implemented but not reported due to journal space constraints. It is also possible that the search strategy missed a number of studies, since there is no current standard for identifying extended follow-up studies in titles or abstracts. Finally, our review only considered studies where the main objective of extended follow-up was to compare the original randomised groups. Extended follow-up may be initiated to answer other types of research questions.

Conclusion

Extended follow-up studies of RCTs can provide vital information about the long-term impacts of an intervention and are an effective use of established trial cohorts. However, the validity of findings from extended follow-up studies relies on appropriate handling of missing data. In this systematic review, we found that a majority of recently published follow-up studies failed to adequately account for missing data in the analysis. This is particularly concerning given the high levels of missing data observed. We encourage researchers working on extended follow-up studies to adhere to recommendations for RCTs by stating the estimand of interest and adopting statistical methods that are valid under a stated assumption about the missing data mechanism. In justifying this assumption, researchers should carefully detail all sources of missing data, including any additional consent processes and eligibility restrictions employed. Sensitivity analyses should also be undertaken to assess the robustness of findings to the assumption made about the missing data in the main analysis. For extended follow-up studies planning an intention to treat analysis, we recommend researchers attempt to collect outcome data on all participants, not just those that adhered to the protocol in the original trial.

7.3.6. Web appendix

Reference list for extended follow-up studies included in systematic review

1. Albada A, van Dulmen S, Spreeuwenberg P and Ausems MG. Follow-up effects of a tailored pre-counseling website with question prompt in breast cancer genetic counseling. *Patient Education and Counseling*. 2015; 98: 69-76.
2. Alehagen U, Aaseth J and Johansson P. Reduced cardiovascular mortality 10 years after supplementation with selenium and coenzyme Q10 for four years: follow-up results of a prospective randomized double-blind placebo-controlled trial in elderly citizens. *PLoS One*. 2015; 10: e0141641.
3. Andersen LL, Ottesen B, Alling Moller LM, et al. Subtotal versus total abdominal hysterectomy: randomized clinical trial with 14-year

- questionnaire follow-up. *American Journal of Obstetrics and Gynecology*. 2015; 212: 758.e1-e54.
4. Aroda VR, Christophi CA, Edelstein SL, et al. The effect of lifestyle intervention and metformin on preventing or delaying diabetes among women with and without gestational diabetes: the Diabetes Prevention Program outcomes study 10-year follow-up. *The Journal of Clinical Endocrinology and Metabolism*. 2015; 100: 1646-53.
 5. Bartelink H, Maingon P, Poortmans P, et al. Whole-breast irradiation with or without a boost for patients treated with breast-conserving surgery for early breast cancer: 20-year follow-up of a randomised phase 3 trial. *Lancet Oncology*. 2015; 16: 47-56.
 6. Bernard K, Hostinar CE and Dozier M. Intervention effects on diurnal cortisol rhythms of Child Protective Services-referred infants in early childhood: preschool follow-up results of a randomized clinical trial. *JAMA Pediatrics*. 2015; 169: 112-9.
 7. Bethoux F, Rogers HL, Nolan KJ, et al. Long-term follow-up to a randomized controlled trial comparing peroneal nerve functional electrical stimulation to an ankle foot orthosis for patients with chronic stroke. *Neurorehabilitation and Neural Repair*. 2015; 29: 911-22.
 8. Blazek S, Rossbach C, Borger MA, et al. Comparison of sirolimus-eluting stenting with minimally invasive bypass surgery for stenosis of the left anterior descending coronary artery: 7-year follow-up of a randomized trial. *JACC Cardiovascular Interventions*. 2015; 8: 30-8.
 9. Burkhard FC, Studer UE and Wuethrich PY. Superior functional outcome after radical cystectomy and orthotopic bladder substitution with restrictive intraoperative fluid management: a followup study of a randomized clinical trial. *The Journal of Urology*. 2015; 193: 173-8.
 10. Collins CT, Gibson RA, Anderson PJ, et al. Neurodevelopmental outcomes at 7 years' corrected age in preterm infants who were fed high-dose docosahexaenoic acid to term equivalent: a follow-up of a randomised controlled trial. *BMJ Open*. 2015; 5: e007314.

11. Courneya KS, Friedenreich CM, Franco-Villalobos C, et al. Effects of supervised exercise on progression-free survival in lymphoma patients: an exploratory follow-up of the HELP Trial. *Cancer Causes Control*. 2015; 26: 269-76.
12. Cuzick J, Sestak I, Cawthorn S, et al. Tamoxifen for prevention of breast cancer: extended long-term follow-up of the IBIS-I breast cancer prevention trial. *Lancet Oncology*. 2015; 16: 67-75.
13. Desai AA, Alemayehu H, Holcomb GW, 3rd and St Peter SD. Minimal vs. maximal esophageal dissection and mobilization during laparoscopic fundoplication: long-term follow-up from a prospective, randomized trial. *Journal of Pediatric Surgery*. 2015; 50: 111-4.
14. Devakumar D, Stocks J, Ayres JG, et al. Effects of antenatal multiple micronutrient supplementation on lung function in mid-childhood: follow-up of a double-blind randomised controlled trial in Nepal. *The European Respiratory Journal*. 2015; 45: 1566-75.
15. Ebenbichler GR, Inschlag S, Pfluger V, et al. Twelve-year follow-up of a randomized controlled trial of comprehensive physiotherapy following disc herniation operation. *Clinical Rehabilitation*. 2015; 29: 548-60.
16. Finfer S, Chittock D, Li Y, et al. Intensive versus conventional glucose control in critically ill patients with traumatic brain injury: long-term follow-up of a subgroup of patients from the NICE-SUGAR study. *Intensive Care Medicine*. 2015; 41: 1037-47.
17. Floege J, Covic AC, Ketteler M, et al. Long-term effects of the iron-based phosphate binder, sucroferric oxyhydroxide, in dialysis patients. *Nephrology, Dialysis, Transplantation*. 2015; 30: 1037-46.
18. Freriks K, Verhaak CM, Sas TC, et al. Long-term effects of oxandrolone treatment in childhood on neurocognition, quality of life and social-emotional functioning in young adults with Turner syndrome. *Hormones and Behavior*. 2015; 69: 59-67.
19. Goicoechea M, Garcia de Vinuesa S, Verdalles U, et al. Allopurinol and progression of CKD and cardiovascular events: long-term follow-up of a

- randomized clinical trial. *American Journal of Kidney Diseases*. 2015; 65: 543-9.
20. Grinstein-Weiss M, Sherraden M, Gale WG, et al. Effects of an individual development account program on retirement saving: follow-up evidence from a randomized Experiment. *Journal of Gerontological Social Work*. 2015; 58: 572-89.
 21. ADAPT-FS Research Group. Follow-up evaluation of cognitive function in the randomized Alzheimer's Disease Anti-inflammatory Prevention Trial and its Follow-up Study. *Alzheimers and Dementia*. 2015; 11: 216-25.e1.
 22. Hahn JY, Yu CW, Park HS, et al. Long-term effects of ischemic postconditioning on clinical outcomes: 1-year follow-up of the POST randomized trial. *American Heart Journal*. 2015; 169: 639-46.
 23. Hassanian-Moghaddam H, Sarjami S, Kolahi AA, Lewin T and Carter G. Postcards in Persia: a twelve to twenty-four month follow-up of a randomized controlled trial for hospital-treated deliberate self-poisoning. *Archives of Suicide Research*. 2015: 1-17.
 24. Hayward RA, Reaven PD, Wiitala WL, et al. Follow-up of glycemic control and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*. 2015; 372: 2197-206.
 25. Hellemans R, Hazzan M, Durand D, et al. Daclizumab versus rabbit antithymocyte globulin in high-risk renal transplants: five-year follow-up of a randomized study. *American Journal of Transplantation*. 2015; 15: 1923-32.
 26. Herold M, Scholz CW, Rothmann F, Hirt C, Lakner V and Naumann R. Long-term follow-up of rituximab plus first-line mitoxantrone, chlorambucil, prednisolone and interferon-alpha as maintenance therapy in follicular lymphoma. *Journal of Cancer Research and Clinical Oncology*. 2015; 141: 1689-95.
 27. Heukelom J, Lopez-Yurda M, Balm AJ, et al. Late follow-up of the randomized radiation and concomitant high-dose intra-arterial or intravenous cisplatin (RADPLAT) trial for advanced head and neck cancer. *Head Neck*. 2015.

28. Hisey MS, Bae HW, Davis RJ, et al. Prospective, randomized comparison of cervical total disk replacement versus anterior cervical fusion: results at 48 months follow-up. *Journal of Spinal Disorders & Techniques*. 2015; 28: E237-43.
29. Humphreys KL, Gleason MM, Drury SS, et al. Effects of institutional rearing and foster care on psychopathology at age 12 years in Romania: follow-up of an open, randomised controlled trial. *Lancet Psychiatry*. 2015; 2: 625-34.
30. Ijas H, Vaarasmaki M, Saarela T, Keravuo R and Raudaskoski T. A follow-up of a randomised study of metformin and insulin in gestational diabetes mellitus: growth and development of the children at the age of 18 months. *BJOG*. 2015; 122: 994-1000.
31. Imamura K, Kawakami N, Furukawa TA, et al. Does Internet-based cognitive behavioral therapy (iCBT) prevent major depressive episode for workers? A 12-month follow-up of a randomized controlled trial. *Psychological Medicine*. 2015; 45: 1907-17.
32. Infante M, Cavuto S, Lutman FR, et al. Long-term follow-up results of the DANTE trial, a randomized study of lung cancer screening with spiral computed tomography. *American Journal of Respiratory and Critical Care Medicine*. 2015; 191: 1166-75.
33. Jafar TH, Jehan I, Liang F, et al. Control of blood pressure and risk attenuation: post trial follow-up of randomized groups. *PLoS One*. 2015; 10: e0140550.
34. Jones NP, Curtis PS and Home PD. Cancer and bone fractures in observational follow-up of the RECORD study. *Acta Diabetologica*. 2015; 52: 539-46.
35. Kalluru R, Ames R, Mason B, et al. Bone density in healthy men after cessation of calcium supplements: 20-month follow-up of a randomized controlled trial. *Osteoporosis International*. 2015; 26: 173-8.
36. Kappos L, O'Connor P, Radue EW, et al. Long-term effects of fingolimod in multiple sclerosis: the randomized FREEDOMS extension trial. *Neurology*. 2015; 84: 1582-91.

37. Karinkanta S, Kannus P, Uusi-Rasi K, Heinonen A and Sievanen H. Combined resistance and balance-jumping exercise reduces older women's injurious falls and fractures: 5-year follow-up study. *Age and Ageing*. 2015; 44: 784-9.
38. Karlsson Videhult F, Ohlund I, Stenlund H, Hernell O and West CE. Probiotics during weaning: a follow-up study on effects on body composition and metabolic markers at school age. *European Journal of Nutrition*. 2015; 54: 355-63.
39. Kenton K, Stoddard AM, Zyczynski H, et al. 5-year longitudinal followup after retropubic and transobturator mid urethral slings. *Journal of Urology*. 2015; 193: 203-10.
40. Khan ZA, Nambiar A, Morley R, Chapple CR, Emery SJ and Lucas MG. Long-term follow-up of a multicentre randomised controlled trial comparing tension-free vaginal tape, xenograft and autologous fascial slings for the treatment of stress urinary incontinence in women. *BJU International*. 2015; 115: 968-77.
41. Kurokawa Y, Sasako M, Sano T, et al. Ten-year follow-up results of a randomized clinical trial comparing left thoracoabdominal and abdominal transhiatal approaches to total gastrectomy for adenocarcinoma of the oesophagogastric junction or gastric cardia. *British Journal of Surgery*. 2015; 102: 341-8.
42. Lee HN, Lee SW, Lee YS, Lee SY and Lee KS. Tension-free vaginal tape-SECUR procedure for the treatment of female stress urinary incontinence: 3-year follow-up results. *Lower Urinary Tract Symptoms*. 2015; 7: 9-16.
43. Lin SY, Tsai CS, Chang YC, et al. The role of pretreatment FDG-PET in treating cervical cancer patients with enlarged pelvic lymph node(s) shown on MRI: a phase 3 randomized trial with long-term follow-up. *International Journal of Radiation Oncology, Biology, Physics*. 2015; 92: 577-85.
44. Liu JT, Li CS, Chang CS and Liao WJ. Long-term follow-up study of osteoporotic vertebral compression fracture treated using balloon kyphoplasty and vertebroplasty. *Journal of Neurosurgery: Spine*. 2015; 23: 94-8.

45. Lorentzen S, Fjeldstad A, Ruud T and Hoglend PA. Comparing short- and long-term group therapy: seven-year follow-up of a randomized clinical trial. *Psychotherapy and Psychosomatics*. 2015; 84: 320-1.
46. Lucas DM, Ruppert AS, Lozanski G, et al. Cytogenetic prioritization with inclusion of molecular markers predicts outcome in previously untreated patients with chronic lymphocytic leukemia treated with fludarabine or fludarabine plus cyclophosphamide: a long-term follow-up study of the US intergroup phase III trial E2997. *Leukemia and Lymphoma*. 2015; 56: 3031-7.
47. Maldonado-Lobon JA, Gil-Campos M, Maldonado J, et al. Long-term safety of early consumption of *Lactobacillus fermentum* CECT5716: A 3-year follow-up of a randomized controlled trial. *Pharmacological Research*. 2015; 95-96: 12-9.
48. Maselko J, Sikander S, Bhalotra S, et al. Effect of an early perinatal depression intervention on long-term child development outcomes: follow-up of the Thinking Healthy Programme randomised controlled trial. *Lancet Psychiatry*. 2015; 2: 609-17.
49. Maynard C, Longstreth WT, Jr., Nichol G, et al. Effect of prehospital induction of mild hypothermia on 3-month neurological status and 1-year survival among adults with cardiac arrest: long-term follow-up of a randomized, clinical trial. *Journal of the American Heart Association*. 2015; 4: e001693.
50. McKelvey L, Schiffman RF, Brophy-Herb HE, et al. Examining long-term effects of an infant mental health home-based early head start program on family strengths and resilience. *Infant Mental Health Journal*. 2015; 36: 353-65.
51. McNamara HC, Wood R, Chalmers J, et al. STOPPIT Baby Follow-up Study: the effect of prophylactic progesterone in twin pregnancy on childhood outcome. *PLoS One*. 2015; 10: e0122341.
52. Molyneux AJ, Birks J, Clarke A, Sneade M and Kerr RS. The durability of endovascular coiling versus neurosurgical clipping of ruptured cerebral

- aneurysms: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT). *Lancet*. 2015; 385: 691-7.
53. Moodie Z, Metch B, Bekker LG, et al. Continued follow-up of Phambili phase 2b randomized HIV-1 vaccine trial participants supports increased HIV-1 acquisition among vaccinated men. *PLoS One*. 2015; 10: e0137666.
 54. Morales A, Espada JP and Orgiles M. A 1-year follow-up evaluation of a sexual-health education program for Spanish adolescents compared with a well-established program. *European Journal of Public Health*. 2016; 26: 35-41.
 55. Naganuma T, Latib A, Sgueglia GA, et al. A 2-year follow-up of a randomized multicenter study comparing a paclitaxel drug-eluting balloon with a paclitaxel-eluting stent in small coronary vessels the BELLO study. *International Journal of Cardiology*. 2015; 184: 17-21.
 56. Ostergaard B, Holbaek E, Sorensen J and Steinbruchel D. Health-related quality of life after off-pump compared with on-pump coronary bypass grafting among elderly high-risk patients: A randomized trial with eight years of follow-up. *European Journal of Cardiovascular Nursing*. 2016; 15: 126-33.
 57. Poulsen SK, Crone C, Astrup A and Larsen TM. Long-term adherence to the New Nordic Diet and the effects on body weight, anthropometry and blood pressure: a 12-month follow-up study. *European Journal of Nutrition*. 2015; 54: 67-76.
 58. Ran MS, Chan CL, Ng SM, Guo LT and Xiang MZ. The effectiveness of psychoeducational family intervention for patients with schizophrenia in a 14-year follow-up study in a Chinese rural area. *Psychological Medicine*. 2015; 45: 2197-204.
 59. Recknor CP, Recker RR, Benson CT, et al. The effect of discontinuing treatment with bloszumab: follow-up results of a phase 2 randomized clinical trial in postmenopausal women with low bone mineral density. *Journal of Bone and Mineral Research*. 2015; 30: 1717-25.
 60. Rohde P, Stice E, Shaw H and Gau JM. Effectiveness trial of an indicated cognitive-behavioral group adolescent depression prevention program

- versus bibliotherapy and brochure control at 1- and 2-year follow-up. *Journal of Consulting and Clinical Psychology*. 2015; 83: 736-47.
61. Ross L, Rottmann N, Andersen KK, Hoybye MT, Johansen C and Dalton SO. Distress after a psychosocial cancer rehabilitation course. Main effects and effect modification in a randomised trial at 12 months of follow-up. *Acta Oncologica*. 2015; 54: 735-42.
 62. Sayal K, Merrell C, Tymms P and Kasim A. Academic outcomes following a school-based RCT for ADHD: 6-year follow-up. *Journal of Attention Disorders*. 2015.
 63. Secher RG, Hjorthoj CR, Austin SF, et al. Ten-year follow-up of the OPUS specialized early intervention trial for patients with a first episode of psychosis. *Schizophrenia Bulletin*. 2015; 41: 617-26.
 64. Sen H, Lam MK, Lowik MM, et al. Clinical events and patient-reported chest pain in all-comers treated with resolute integrity and promus element stents: 2-year follow-up of the DUTCH PEERS (DURable Polymer-Based STent CHallenge of Promus ElemEnt Versus ReSolute Integrity) randomized trial (TWENTE II). *JACC Cardiovascular Interventions*. 2015; 8: 889-99.
 65. Sparano JA, Zhao F, Martino S, et al. Long-term follow-up of the E1199 phase III trial evaluating the role of taxane and schedule in operable breast cancer. *Journal of Clinical Oncology*. 2015; 33: 2353-60.
 66. Stagl JM, Bouchard LC, Lechner SC, et al. Long-term psychological benefits of cognitive-behavioral stress management for women with breast cancer: 11-year follow-up of a randomized controlled trial. *Cancer*. 2015; 121: 1873-81.
 67. Stice E, Rohde P, Butryn ML, Shaw H and Marti CN. Effectiveness trial of a selective dissonance-based eating disorder prevention program with female college students: Effects at 2- and 3-year follow-up. *Behaviour Research and Therapy*. 2015; 71: 20-6.
 68. Svege I, Nordsletten L, Fernandes L and Risberg MA. Exercise therapy may postpone total hip replacement surgery in patients with hip osteoarthritis: a

- long-term follow-up of a randomised trial. *Annals of the Rheumatic Diseases*. 2015; 74: 164-9.
69. Tamirou F, D'Cruz D, Sangle S, et al. Long-term follow-up of the MAINTAIN Nephritis Trial, comparing azathioprine and mycophenolate mofetil as maintenance therapy of lupus nephritis. *Annals of the Rheumatic Diseases*. 2016; 75: 526-31.
 70. Tepe G, Schnorr B, Albrecht T, et al. Angioplasty of femoral-popliteal arteries with drug-coated balloons: 5-year follow-up of the THUNDER trial. *JACC Cardiovascular interventions*. 2015; 8: 102-8.
 71. Thomsen FB, Brasso K, Christensen IJ, et al. Survival benefit of early androgen receptor inhibitor therapy in locally advanced prostate cancer: long-term follow-up of the SPCG-6 study. *European Journal of Cancer*. 2015; 51: 1283-92.
 72. Tommaselli GA, D'Afiero A, Di Carlo C, Formisano C, Fabozzi A and Nappi C. Tension-free vaginal tape-obturator and tension-free vaginal tape-Secur for the treatment of stress urinary incontinence: a 5-year follow-up randomized study. *European Journal of Obstetrics, Gynecology and Reproductive Biology*. 2015; 185: 151-5.
 73. Trilok-Kumar G, Kaur M, Rehman AM, et al. Effects of vitamin D supplementation in infancy on growth, bone parameters, body composition and gross motor development at age 3-6 years: follow-up of a randomized controlled trial. *International Journal of Epidemiology*. 2015; 44: 894-905.
 74. Valkenburg AJ, van den Bosch GE, de Graaf J, et al. Long-term effects of neonatal morphine infusion on pain sensitivity: follow-up of a randomized controlled trial. *Journal of Pain*. 2015; 16: 926-33.
 75. van Gent WB, Catarinella FS, Lam YL, et al. Conservative versus surgical treatment of venous leg ulcers: 10-year follow up of a randomized, multicenter trial. *Phlebology*. 2015; 30: 35-41.
 76. Veronesi G, Lazzeroni M, Szabo E, et al. Long-term effects of inhaled budesonide on screening-detected lung nodules. *Annals of Oncology*. 2015; 26: 1025-30.

77. Webb NJ, Douglas SE, Rajai A, et al. Corticosteroid-free kidney transplantation improves growth: 2-year follow-up of the TWIST randomized controlled trial. *Transplantation*. 2015; 99: 1178-85.
78. Wen LM, Baur LA, Simpson JM, et al. Sustainability of effects of an early childhood obesity prevention trial over time: a further 3-year follow-up of the Healthy Beginnings trial. *JAMA Pediatrics*. 2015; 169: 543-51.
79. Williksen JH, Husby T, Hellund JC, Kvernmo HD, Rosales C and Frihagen F. External fixation and adjuvant pins versus volar locking plate fixation in unstable distal radius fractures: a randomized, controlled study with a 5-year follow-up. *Journal of Hand Surgery (American Volume)*. 2015; 40: 1333-40.
80. Zhang M, Li Q, Tie HT, Jiang YJ and Wu QC. Methods of reconstruction after esophagectomy on long-term health-related quality of life: a prospective, randomized study of 5-year follow-up. *Medical Oncology*. 2015; 32: 122.
81. Zupi E, Centini G, Lazzeri L, et al. Hysteroscopic endometrial resection versus laparoscopic supracervical hysterectomy for abnormal uterine bleeding: long-term follow-up of a randomized trial. *Journal of Minimally Invasive Gynecology*. 2015; 22: 841-5.

*** End of published article ***

7.4. Guidance on the use of multiple imputation

The systematic review presented in Section 7.3 provides an overview of how missing outcome data are handled in published extended follow-up studies. A further aim of this thesis is to provide guidance on the implementation of MI in this setting. Although MI was rarely used in the extended follow-up studies included in the systematic review, valuable information on the extent and common sources of missing outcome data in this setting was obtained. In particular, the systematic review indicated that the amount of missing data in extended follow-up studies tends to be high and that eligibility restrictions and

separate consent processes are often used. The implications of eligibility restrictions and separate consent processes on the implementation of MI are now considered in more detail.

It is assumed throughout this section of the thesis that the goal of analysis is to provide an unbiased and efficient estimate of the estimand of interest in the extended follow-up study, hereafter referred to simply as the estimand. Importantly, to focus on issues in implementing MI, the appropriateness or otherwise of specific estimands is not evaluated in this section. Suppose for example the estimand relates to the effect of treatment in participants that complied with their allocated intervention in the original trial. Since randomised groups are unlikely to be comparable once non-compliant participants are excluded from consideration, such an estimand provides a measure of association rather than causation. Rather than questioning the merits of the estimand, we consider only the implementation of MI for best estimating it. Attention is also restricted to settings where the design of the extended follow-up is consistent with its estimand. Most notably, we do not tackle the case in which an eligibility restriction based on adherence to the protocol in the original trial is employed, yet where interest in the extended follow-up study concerns the ITT estimand.

7.4.1. Multiple imputation and eligibility restrictions

Before detailing common eligibility restrictions in extended follow-up studies and their implications for the use of MI in this setting, it is useful to clarify what was intended by the term “eligibility restriction” in the systematic review. Broadly, an eligibility restriction was taken to be any rule that prevented individuals who otherwise could have taken part in the extended follow-up from participating in this phase of the trial. In addition to eligibility restrictions, individuals could be precluded from participating in extended follow-up for the following reasons:

- death;
- loss to follow-up;

- withdrawal from the original RCT preventing further contact;
- non-selection in a random sample chosen for extended follow-up; and
- non-consent to extended follow-up.

While some studies included in the systematic review stated that loss to follow-up or non-consent to extended follow-up rendered a participant ineligible, for consistency in this discussion the reason for not participating in extended follow-up in these cases was taken to be loss to follow-up or non-consent, respectively (rather than ineligibility). Conversely, withdrawal from the original RCT was classified as an eligibility restriction if a study both described it as such, and did not explicitly detail whether the withdrawal process prevented further contact with participants.

Of the 81 articles included in the review, 36 (44%) reported eligibility restrictions for entry into extended follow-up. The most common class of eligibility restriction concerned adherence to the protocol in the original RCT (22 studies), typically defined according to satisfactory completion of outcome assessments and/or sufficient compliance with the allocated intervention. Studies where withdrawal from the original RCT was taken to be an eligibility restriction according to the criteria given above were also included in this eligibility restriction class, as withdrawing from a study generally entails incomplete outcome assessments and/or non-compliance with the intervention. Studies in the systematic review also ruled participants ineligible according to their enrolling centre (three studies), randomised arm (two studies, both involving three arms in the original RCT), geographic availability (three studies), or other baseline (three studies) or post-randomisation (three studies) characteristics. Across studies reporting eligibility restrictions, the median percentage of participants randomised in the original trial that were eligible for extended follow-up was 86%.

An important function of eligibility restrictions, as identified in the systematic review, is to limit entry into the extended follow-up to participants contained within the target population of the estimand. Should the intention of the extended

follow-up be to estimate the effect of treatment in individuals who complied with their allocated intervention, for example, then it is logical to restrict participation to compliers. Likewise, if the goal of extended follow-up is to estimate the effect of treatment in participants with particular baseline characteristics, only participants with these characteristics need to be included in the follow-up study. Using the Normoglycemia in Intensive Care Evaluation-Survival Using Glucose Algorithm Regulation (NICE-SUGAR) study as a specific example, excluding participants without traumatic brain injury at baseline was consistent with the aim of this extended follow-up study to “compare the effect of intensive versus conventional blood glucose control in patients with traumatic brain injury” (139). Another function of eligibility restrictions is to ensure the logistical feasibility of successfully completing the extended follow-up study. For example, in multicentre trials it may only be feasible to recruit participants randomised at some of the centres (e.g. (141, 142)), or, for international trials, to centres within particular countries (e.g. (143)). Alternatively, participation could be restricted to individuals living within reasonable geographic proximity of the research team at the time of the extended follow-up (e.g. (144-146)).

Although not observed in the systematic review, it is also conceivable that eligibility restrictions could be employed based on statistical power considerations. In particular, if the sample size required to achieve the desired power for the primary outcome of the extended follow-up study is substantially less than the sample size of the original trial, then an additional eligibility restriction could be applied to reduce the sample size. For an estimand defined for all randomised individuals, ideally the chosen eligibility restriction would not lead to systematic differences between eligible and ineligible participants. Candidate eligibility restrictions in this case could be based on enrolling centre (if eligible centres are considered representative of all participating centres) or the chronological order of participants in the randomisation sequence (e.g. restricting entry to the first 100 participants randomised, assuming that the characteristics of participants did not change over time). Rather than applying an eligibility restriction, where the ability to participate in extended follow-up is decided

according to a deterministic rule, the reduced sample size might instead be achieved by randomly selecting a sample of participants to take part in extended follow-up. This latter situation is slightly different to an eligibility restriction as each participant has some chance of taking part in the extended follow-up study, which has implications for analysis. This situation is considered separately in Section 7.4.3.

An important consideration when implementing MI in extended follow-up studies with eligibility restrictions is whether participants deemed ineligible for extended follow-up should be included in the imputation model. In practice, a useful way to approach this question is to first identify whether ineligible participants are contained within the target population of the estimand. Importantly, if there is no interest in the effect of treatment in participants ruled ineligible, then clearly there is no need to impute missing outcome data in these participants. Again using the extended follow-up of the NICE-SUGAR study as an example (139), there is no reason to impute missing outcome data in participants ruled ineligible due to not having a traumatic brain injury at baseline, if interest lies only in the effect of treatment in those with traumatic brain injury. A general recommendation then is that ineligible participants should not be included in the imputation model when the function of the eligibility restriction is to limit participation to individuals contained within the target population of the estimand.

In settings where eligibility restrictions are used to ensure the feasibility of the extended follow-up study or to reduce the sample size according to a power calculation, it may be the case that participants ruled ineligible are contained within the target population of the estimand. Indeed, the systematic review identified several extended follow-up studies that employed eligibility restrictions for the sake of feasibility while also reporting interest in the ITT estimand (i.e. the effect of randomisation over all randomised individuals). In studies such as these, a general recommendation is to include ineligible participants in the imputation model if this is likely to lead to improved estimation of the estimand. As well as the ability to satisfy an assumption about the missing data mechanism, the

decision of whether or not to include ineligible participants in the imputation model could be influenced by the availability of auxiliary variables in the original RCT. Importantly, if there is little auxiliary information to aid with the imputation of outcomes collected during extended follow-up, then including ineligible participants in the imputation model may simply add noise to the estimation process.

Clearly eligibility restrictions could be employed for reasons other than those identified in the systematic review, and so the above recommendations are limited to the scenarios encountered in this review. It is also possible that multiple eligibility restrictions could be applied within a single extended follow-up study, in which case it may be reasonable to impute missing outcome data for some ineligible participants and not others. In light of this, ultimately the choice of whether to impute missing outcome data for ineligible participants is perhaps best evaluated on a case by case basis, with careful justification for the decision made.

7.4.2. Multiple imputation and separate consent processes

Depending on the information provided to participants in the original RCT and the specifics of the extended follow-up study in question, it may be necessary to obtain informed consent from participants prior to initiating the extended follow-up (16). In the systematic review, 24 of the 81 included studies (30%) reported the use of a separate consent process for entry into the extended follow-up. The median percentage of participants randomised in the original trial providing consent in these studies was 70%. It is possible, of course, that additional studies employed separate consent processes for extended follow-up but failed to report them.

A key consideration when applying MI in extended follow-up studies with separate consent processes is whether missing outcome data should be imputed for participants who were approached but failed to provide consent to the extended follow-up. Arguably this problem is more straightforward than in the

corresponding case for eligibility restrictions. Unlike eligibility restrictions, which are often applied to limit recruitment to those participants contained in the target population of the estimand, separate consent processes function only to educate individuals about study processes so they can make an informed decision about participating. As a result, there is typically interest in the effect of treatment in non-consenting participants. Indeed, of those studies included in the systematic review that explicitly defined the estimand and reported the use of a separate consent process (11 studies), not one excluded non-consenting participants from the target population of the estimand.

Assuming the goal of the analysis is to estimate the effect of treatment in a population that includes non-consenting participants, the decision of whether to include non-consenting participants in the imputation model can be based on the expected bias and efficiency of the treatment effect estimate. Factors that could influence this decision once again include the availability of auxiliary variables to assist with the imputation of missing outcomes, and the ability to satisfy an assumption about the missing data mechanism. Ultimately it is recommended that non-consenting participants be included in the imputation model if this is likely to lead to improved estimation of the estimand.

7.4.3. Multiple imputation and other sources of missing data

As described previously, individuals could be precluded from participating in extended follow-up due to death, loss to follow-up, withdrawal from the original RCT preventing further contact, or non-selection in a random sample chosen for extended follow-up. Participants recruited into the extended follow-up phase of the trial could also fail to contribute outcome data. With the exception of non-selection in a random sample, these potential sources of missing outcome data are common to standard RCTs, hence recommendations for their handling in this context can also be applied to extended follow-up studies. In particular, missing outcome data should be imputed in an analysis involving MI for participants contained within the target population of the estimand (1). As with ineligible or

non-consenting participants, one difficulty in imputing missing outcome data in participants who withdraw or are lost to follow-up before the commencement of extended follow-up is that there will be no information collected during the extended follow-up to aid imputation. Although ideally these participants would be included in the imputation model, the lack of auxiliary information to impute the missing values may simply add noise to the estimation procedure. In the case of non-selection in a random sample chosen for extended follow-up, the probability of selection will be known to researchers, hence a weighting approach might be applied in place of MI to handle missing outcome data.

Although not a focus of this thesis, the issue of how to handle unobserved outcome data due to participant death warrants brief mention here. Importantly, if an outcome variable is not considered meaningful in participants that died, then the outcome should be considered undefined rather than missing in these participants and MI should not be applied (1). Several statistical approaches have been proposed to address undefined outcome data due to death. If death is known to be unrelated to treatment, the effect of treatment can simply be estimated using data from surviving patients (1). Could death be related to treatment, it may be possible to include death as a component of a composite outcome, or to attribute to death a utility score on the same scale as the outcome (66). Alternatively, principal stratification could be used to estimate the effect of treatment in the subset of participants who would have remained alive on either treatment (147). For further discussion on these and other techniques for addressing undefined outcome data due to death, see (1, 66).

7.4.4. Inverse probability weighting and multiple imputation

Assuming interest in the effect of treatment over all randomised participants, as with an ITT analysis, a concern with implementing MI is that a substantial proportion of randomised participants could be missing data on all variables collected during the extended follow-up. Since the imputation model describes the joint distribution of all variables subject to missing data, a greater number of

variables requiring imputation is likely to mean an increased risk of imputation model misspecification. In turn, any deficiencies in the imputation model will have a proportionately larger effect on estimation when larger amounts of data require imputation. Ultimately this could lead to serious bias. Another concern is that participants with missing data on a large number of variables might simply add noise to the imputation process, which could result in reduced efficiency.

In settings where data are missing on many variables in many participants, as is often the case in extended follow-up studies, IPW may be an appealing alternative to MI. Whereas MI requires appropriate specification of a joint (i.e. multivariate) model for the missing data conditional on the observed data in order to produce valid inference, IPW only requires an appropriately specified univariate model for the probability that an individual has complete data. Of course, as with any approach to handling missing data, IPW is not without limitations. In its standard implementation, IPW can be inefficient relative to MI, as it discards information from partially observed cases. Further, the approach can be difficult to implement for non-monotone patterns of missing data (26).

Another possibility for handling missing outcome data, as introduced in Seaman et al. (148), is to combine IPW and MI. The basic idea of this approach, termed “IPW/MI”, is to use MI to account for missingness in participants with few variables subject to missing data, and IPW to account for participants with larger blocks of missing data. In this way, IPW/MI could acquire some of the efficiency advantages of MI while minimising potential bias due to imputing large blocks of missing data. A key precursor in applying IPW/MI is the determination of a rule for when to include a participant in the imputation model. In the context of extended follow-up studies, a sensible rule might be to impute results for individuals who participate in extended follow-up but have sporadic missing data in outcomes collected during this phase of the trial, and to use IPW to handle individuals who did not participate in the extended follow-up. After defining the inclusion rule for imputation, missing values in included participants are multiply imputed using standard MI techniques. Resulting completed datasets are then

analysed separately using IPW, that is, with included participants weighted in the analysis according to the inverse of the probability of satisfying the inclusion rule, and with a robust error variance calculated to account for the weights. Finally, results from the weighted analyses are then combined using Rubin's rules, which have been shown to perform well following IPW/MI (see Seaman et al. (148) for details).

Arguably the IPW/MI approach is best reserved for extended follow-up studies where data are missing on many variables in many participants, and where there is concern over the appropriate specification of the imputation model. In these settings, IPW/MI could be employed as a primary method of analysis, as a form of sensitivity analysis, or as a diagnostic check for MI. Should IPW/MI and standard MI produce similar results, this might offer reassurance that the imputation model is appropriately specified. Conversely, should the results of IPW/MI and standard MI differ greatly, this might highlight ways in which the imputation model could be improved.

7.4.5. Case study

To illustrate some of the challenges in implementing MI in extended follow-up studies, and how IPW/MI can be used as a diagnostic check for MI in this setting, once again the DINO trial was considered (91). As described previously, in DINO $n = 657$ preterm infants born < 33 weeks gestation were randomised to receive a high docosahexaenoic acid (DHA) or a standard DHA diet from within 5 days of commencing enteral feeds through to term-equivalent age. Randomisation was stratified by centre (5 centres), sex, and birth weight ($< 1250\text{g}$, $\geq 1250\text{g}$), with infants from a multiple birth randomised according to the sex and birth weight of the first born infant. The initial DINO trial concluded following the assessment of neurodevelopmental outcomes in the children at 18 months corrected age. Later an extended follow-up period was initiated to assess neurodevelopmental and growth outcomes in the children at 7 years corrected age. Consent to participate in the extended follow-up phase of the trial was obtained from a parent or guardian

prior to the initiation of extended follow-up. No eligibility restrictions were employed for entry into extended follow-up. To illustrate the application of MI to this study, once again the primary ITT analysis of fat free mass (FFM) at 7 years corrected age was considered. To simplify the dataset for illustration purposes, second and subsequent born infants from a multiple birth and infants that died before the extended follow-up were ignored, resulting in an example dataset with 262 and 258 infants in the high and standard DHA groups, respectively.

The flow of children through the original DINO trial and its extended follow-up phase is summarised in Table 7.5. As shown in this table, only 25 of the 520 randomised children (4.8%) failed to enter the extended follow-up phase of the trial; 9 were lost to follow-up, 9 were withdrawn during the original trial and could not be re-contacted, and 7 had families that were approached but failed to consent to extended follow-up. This remarkably high retention rate was attributed partly to the families of the preterm children, who were keen for their child's development to be monitored, and partly to the efforts of the research team in keeping in regular contact with families prior to the commencement of extended follow-up. Of the 495 children who entered the extended follow-up phase of DINO, 10 later withdrew consent and a further 11 were unable to secure an appointment for an outcome assessment (and hence had no outcome data in the extended follow-up study).

Table 7.5. Flow of children through the DINO extended follow-up study^a.

Group	High DHA	Standard DHA	Total
Number of children randomised in original trial	262	258	520
Number of children not entering extended follow-up phase	16	9	25
Loss to follow-up	5	4	9
Withdrawal from original RCT preventing further contact	6	3	9
Family approached but did not consent to extended follow-up	5	2	7
Number of children whose families consented to extended follow-up	246	249	495
Number of children providing some outcome data during extended follow-up	237	237	474
Number of children who did not complete any assessments during extended follow-up	9	12	21
Withdrawal during extended follow-up	5	5	10
Unable to secure appointment	4	7	11

^a Numbers exclude second and subsequent born infants from a multiple birth and infants that died before the commencement of the extended follow-up study

As described previously (see Section 6.3.8), FFM was missing for 65/262 (24.8%) and 46/258 (17.8%) children in the high and standard DHA groups, respectively. Key predictors of FFM that could potentially be useful for imputation included centre, sex, and weight, height and systolic blood pressure at 7 years corrected age. Since centre and sex were baseline measures, these variables were treated as covariates for adjustment in the analysis models. In contrast, the post-randomisation measures of weight, height, and systolic blood pressure at 7 years corrected age were treated as auxiliary variables in analyses involving MI.

Table 7.6 summarises the patterns of missing data on FFM and the three auxiliary variables used for imputation. As displayed in this table, 403 of the 520 randomised children (77.5%) provided complete data on all of these variables. Conversely, 53/520 children (10.2%) failed to contribute any data on these extended follow-up measures. Another common pattern was to have complete data on all variables except for the outcome variable FFM (pattern 3, n = 35).

Table 7.6. Missing data patterns for fat free mass and key auxiliary variables.

Pattern	Frequency (%)	Fat free mass	Weight	Height	Systolic blood pressure
1	403 (77.5)	+	+	+	+
2	6 (1.2)	+	+	+	-
3	35 (6.7)	-	+	+	+
4	11 (2.1)	-	+	+	-
5	1 (0.2)	-	+	-	+
6	11 (2.1)	-	-	+	+
7	53 (10.2)	-	-	-	-

+ indicates observed data, - indicates missing data

In estimating the effect of treatment on FFM, the following three MI strategies were considered:

1. MI including all n = 520 randomised children;
2. MI restricted to the n = 467 children who contributed at least some data on FFM and the three auxiliary variables; and

3. IPW/MI, with MI applied to the $n = 467$ children who contributed at least some data on FFM and the three auxiliary variables, and IPW used to recover the sample size to the 520 randomised children.

For each MI strategy, missing values in FFM and the three auxiliary variables were imputed using a Markov chain Monte Carlo algorithm assuming multivariate normality (11). Each imputation model involved a burn-in of 5000 iterations, with $m = 50$ complete datasets created. Imputation was performed separately by randomised group according to the findings of Chapter 6, with the fully observed baseline covariates centre and sex also added to imputation models to ensure consistency with the intended analysis. For IPW/MI, logistic regression analysis revealed that the odds of failing to contribute any data during extended follow-up was higher in one of the five study centres (odds ratio vs. coordinating centre = 2.37; 95% CI 1.08 to 5.21; $p = 0.03$) and decreased with the age of the child's mother at randomisation (odds ratio = 0.92; 95% CI 0.88 to 0.97; $p = 0.001$). As centre and mother's age at randomisation were both fully observed, weights were calculated directly from a logistic regression model involving these two predictors. It is worth noting that weights can also be calculated using incomplete predictors of missing data, although the statistical procedure is more complex than with complete predictors (see (26) for details).

Treatment effect estimates from the three methods for handling missing outcome data are presented in Table 7.7. Results from a complete case analysis, both unadjusted and adjusted for centre and sex, are also presented for comparison. As shown in the table, treatment effect estimates and corresponding 95% confidence intervals were very similar for the three MI approaches. Evidently the decision of whether to incorporate the 53 children with missing data on FFM and the three auxiliary variables made little difference to estimation. This finding might be attributable both to the small number of children accounted for in this group (10.2% of the randomised sample), and the small amount of information on the effect of treatment provided by these children. Perhaps these children might have contributed more information about the treatment effect had useful auxiliary

variables for FFM been available from the original DINO trial. Comparing MI to all randomised children (n = 520) with IPW/MI, the similar estimates of the treatment effect suggests that any bias due to imputing large blocks of missing data was likely minimal. This demonstrates the usefulness of IPW/MI as a diagnostic check for MI.

Table 7.7. Treatment effect estimates for fat free mass (kg) at 7 years corrected age from the DINO extended follow-up study.

Method of analysis	Mean difference	Standard error	95% confidence interval
Unadjusted CCA (n = 409)	-0.007	0.259	-0.514 to 0.500
CCA (n = 409)	0.048	0.238	-0.420 to 0.515
MI to full randomised group (n = 520)	-0.118	0.227	-0.563 to 0.327
MI to those with some data at follow-up (n = 467)	-0.104	0.229	-0.553 to 0.346
MI to those with some data at follow-up (n = 467) + IPW (n = 53)	-0.108	0.230	-0.559 to 0.344

Abbreviations: CCA, complete case analysis; MI, multiple imputation; IPW, inverse probability weighting.

Although the treatment effect estimate was stable across the different MI strategies in this example, such a pattern of results might not be seen in other extended follow-up studies. Importantly, unlike many of the extended follow-up studies included in the systematic review, DINO did not involve eligibility restrictions for participation in the extended follow-up. In addition, the overall percentage of randomised participants with complete data on FFM was 78.7%, quite a bit higher than the median value of 65.9% observed in the systematic review. Finally, the imputation model employed in this case study was relatively simple, involving just the four variables with missing data, and so was unlikely to be substantially misspecified. All of these factors may have contributed to the similar performance of the MI strategies in the current example.

7.4.6. Conclusions

When implementing MI to handle missing outcome data in extended follow-up studies, a key task is to identify which participants are contained in the target population for the estimand. Importantly, if there is no interest in the effect of treatment in participants with particular characteristics, then there is no need to

include these participants in the imputation model. Often participants ruled ineligible for extended follow-up will not be of interest for analysis, while participants that have missing outcome data for other reasons will be. Having established the group of participants of interest for analysis, it may be the case that a large proportion have missing data on a range of variables due to not partaking in the extended follow-up. In this case, it can be useful to contrast the results of a standard MI analysis (i.e. involving all participants in the target population) with an MI analysis restricted to those participants who commenced extended follow-up, and/or with IPW/MI. Should these approaches produce similar results, this would offer reassurance that results are robust to the decision regarding the handling of the missing data. Conversely, differences between the approaches would highlight the sensitivity of results to the assumption made about the missing data, and potentially suggest ways in which the imputation model might be refined. As with any other research setting subject to missing data, additional sensitivity analyses in which the assumption about the missing data mechanism is relaxed should also be undertaken.

8. Summary and conclusions

This thesis has explored several issues in the practical application of MI for handling missing outcome data. In particular, the thesis has addressed specific aims concerning the imputation of missing outcome data (1) in the presence of auxiliary variables, (2) for estimating relative risks, (3) in RCTs, and (4) in extended follow-up studies based on RCTs. New contributions from this thesis to the field are timely given the increasing popularity of MI and the widespread occurrence of missing outcome data in the medical literature. In this final chapter, key findings and contributions are summarised, limitations of the work are discussed, and suggestions for further research are highlighted.

8.1. Key findings and contributions

8.1.1. Thesis aim 1

The first aim of this thesis was to compare the performance of MI and MID in settings where missing data are evident in both outcome and exposure variables, and where auxiliary variables associated with the outcome are included in the imputation model. Two types of auxiliary variables were of interest in this investigation: those associated just with the outcome, which would be included for efficiency gains, and those associated with both the outcome and missingness in the outcome, which would be included for efficiency gains and bias reduction.

As described in Chapter 4, the performance of MI and MID in the presence of an auxiliary variable for the outcome was evaluated using data simulation. In simulation settings where the auxiliary variable was associated with the outcome, but not missingness in the outcome, both MI and MID exhibited negligible bias in estimating regression parameters when data were MAR. In terms of precision, MID performed marginally better than standard MI when there was a weak correlation between the auxiliary variable and the outcome, while MI was noticeably more efficient than MID for moderate-to-strong correlations. In

simulation settings where the auxiliary variable was associated with both the outcome and missingness in the outcome, it was shown for the first time that MID produces biased estimates of regression parameters when data are MAR, whereas standard MI does not. The magnitude of the bias with MID increased with the amount of missing data and with the strength of the correlation between the auxiliary variable and the outcome.

The practical implications of this research are that if the imputation model includes auxiliary variables for the outcome, then it is important that imputed outcomes are kept in the analysis. MID is better reserved for settings where auxiliary variables for the outcome are unavailable.

8.1.2. Thesis aim 2

The second major aim of this thesis was to assess the performance of standard model-based methods of MI for handling missing data in outcome and exposure variables when estimating the relative risk. While relative risks are typically estimated using log binomial models, standard model-based methods for imputing incomplete binary outcomes involve logistic regression or an assumption of multivariate normality. It was unclear whether inconsistencies between imputation and analysis models in this setting could result in biased and/or inefficient estimates of the relative risk. A supplementary aim was to evaluate whether deleting imputed outcomes prior to analysis improves the performance of MI in this setting.

The performance of standard model-based methods of MI for handling missing data when estimating the relative risk was evaluated in Chapter 5 using data simulation. The investigation considered the performance of MVNI and FCS with a logistic imputation model for the outcome, with both MI approaches applied with or without the deletion of imputed outcomes prior to analysis. Results indicated that MVNI is likely to be a poor choice for handling missing data when interest concerns the relative risk, with the approach consistently producing

estimates of the relative risk that were biased towards the null. Deleting imputed outcomes following MVNI tended to reduce the bias of this imputation method, but this came at the expense of decreased efficiency. Although outperforming MVNI, FCS was also associated with biased estimates of the relative risk, with the magnitude of the bias positively associated with the outcome prevalence and the size of the relative risk. Deleting imputed outcomes following FCS did not improve the performance of this imputation approach.

As the first study to explore the performance of standard model-based methods of MI for estimating the relative risk, this work has important practical implications. Most notably, the research shows that FCS with a logistic imputation model for the outcome, despite its shortcomings, should be preferred over MVNI for handling an arbitrary pattern of missing data in outcome and exposure variables when estimating the relative risk. Further, imputed outcomes should be retained for analysis in this setting. In demonstrating performance deficits with both MVNI and FCS when estimating the relative risk, these findings reinforce the importance of appropriately replicating the functional form of a chosen analysis within the imputation model. Ultimately, it is hoped that this research will lead to the development of new approaches within the MI framework for more suitably handling missing outcome data when estimating the relative risk.

8.1.3. Thesis aim 3

The third aim of this thesis was to evaluate the performance of MI for handling missing outcome data in RCTs, and to explore the merits of imputing overall and separately by randomised group in this context. Of interest was the utility of MI for estimating treatment effects according to the ITT principle. There were two primary motivating reasons for undertaking this work. First, editors and journal reviewers are increasingly requesting the use of MI to handle missing outcome data in RCTs, despite limited evidence that MI outperforms alternative statistical approaches in this setting. Second, MI is often implemented separately by randomised group in RCTs in order to facilitate subgroup analyses, however

whether this approach might also offer benefits in settings where subgroup analyses are not of interest had not been previously investigated.

In line with theoretical results in the literature, MI was observed to produce unbiased treatment effect estimates in simulation settings where outcome data were MAR and where imputation and analysis models were correctly specified (see Chapter 6). However, MI was often less efficient than alternative unbiased approaches for handling missing data in RCTs. For example, MI was less efficient than a CCA for univariate outcomes with missing data and the likelihood-based LMM for continuous multivariate outcomes with missing data. In settings where the analysis model overlooked an interaction effect involving randomised group, MI only produced unbiased estimates of the average treatment effect when implemented separately by randomised group.

A key contribution of this research to the literature is that it demonstrates that MI should never be seen as the only acceptable option for handling missing outcome data in RCTs. In many cases a simpler approach to missing data can be preferable. The work also indicates that where MI is employed in the analysis of an RCT, imputation should be performed separately by randomised group. Compared to including all randomised participants in a single imputation model, imputing separately by randomised group offers greater robustness against imputation model misspecification at little cost. It is hoped that the publication from this work will be a useful reference for researchers involved with the analysis of RCT data, and for editors and journal reviewers tasked with judging the appropriateness of statistical methods for handling missing outcome data in reports of RCTs.

8.1.4. Thesis aim 4

The fourth and final aim of this thesis was to review the extent and common sources of missing outcome data in recently published extended follow-up studies, and to provide general recommendations around the implementation of MI in this setting. This aim was developed in response to the potentially serious threat to

inference posed by missing outcome data in extended follow-up studies, and to the scarcity of literature on this type of study design.

As described in Chapter 7 of the thesis, a systematic review of recently published extended follow-up studies was undertaken to characterise the nature and handling of missing outcome data in this setting. High rates of missing outcome data were observed in the review, an unsurprising finding given that primary outcomes in included studies were collected at a median of 7 years after randomisation in the original trial. As well as attrition over time, eligibility restrictions and consent processes for entry into extended follow-up were common reasons why randomised participants failed to contribute outcome data during this phase of the trial. Despite the serious threat to inference presented by missing outcome data, the statistical approaches used to address this problem in the studies reviewed were often inadequate. Importantly, only half of the included studies defined the estimand of interest, less than 10% stated the missing data mechanism assumed in the analysis, and just 25% undertook sensitivity analyses around the missing data mechanism. In addition, more than half the included studies performed the main analysis under the strong and often unrealistic assumption that outcome data were MCAR.

Findings from the systematic review were used to develop recommendations around the implementation of MI as a primary method of analysis in extended follow-up studies. The main recommendations were to include participants in the imputation model when (a) they were of interest for analysis, and (b) where their inclusion would likely lead to improved estimation of the chosen estimand. It was also suggested that IPW/MI and/or MI restricted to participants who commenced extended follow-up could be used as a form of sensitivity analysis or diagnostic check for a standard MI analysis involving all participants of interest.

As the first study to quantify the considerable threat posed by missing outcome data in extended follow-up studies, it is hoped that this research will raise awareness of this problem and lead to the adoption of more suitable statistical

approaches when analysing such studies. It is also hoped that recommendations from this research will simplify the process of applying MI in the analysis of extended follow-up studies, particularly around how to address missing outcome data resulting from the use of eligibility restrictions and separate consent processes.

8.2. Limitations and future directions

The limitations of each individual study contributing to this thesis have been described in the relevant chapter discussions. In this section, the limitations of the thesis as an overall body of work are discussed, and areas for future research are identified.

This thesis has relied heavily on data simulation to evaluate the performance of MI for handling missing outcome data. In order to attribute any deficiencies in performance to the method of MI, only simple simulation scenarios were considered throughout the thesis. In particular, attention was restricted to main effects models involving at most two covariates, where individual variables in the analysis model followed either a normal or a Bernoulli distribution. In practice, interest might concern more complex relationships (e.g. containing interaction terms) involving a larger number of variables from a variety of different distribution types. Although similar performance might be anticipated with MI in more complex practical settings, additional simulation studies are needed to determine whether the findings of this thesis extend to such settings. In addition to the focus on simple analysis models, this thesis considered only a narrow assortment of missing data mechanisms. In particular, data in outcome and exposure variables were set to be missing according to simple logistic regression models or, following previous simulation work on the MID method, according to the cumulative distribution function of the normal distribution. In settings where bias and precision losses were evident with a given method of MI, it is possible that performance deficits could be quite sensitive to the functional form of the

MAR mechanism. Hence it would be beneficial for future research to expand upon the range of missing data mechanisms considered.

Another limitation of this thesis is that it only considered the application of MI under an assumption that data were MAR. Although this corresponds with the standard implementation of MI, in any given analysis data may instead be MNAR, which occurs when the probability of missing data depends on unobserved values. Although the implementation of MI under an MNAR assumption is an active area of research, there is little indication that this research extends to problems such as the estimation of relative risks or the handling of missing outcome data in extended follow-up studies. Given the utility of MI under an MNAR assumption, this is an area for future research.

This thesis focused predominantly on the application of MI in settings where observations were independent. Yet many datasets in medical research involve some form of clustering, where observations can be classified into a number of distinct groups or “clusters”, such that observations within the same cluster are likely to be more similar than observations in different clusters. Common examples include repeated measurements on the same participant over time (i.e. longitudinal data), students within schools, or studies within a meta-analysis. Another example of clustering is provided by the DINO case study, where infants from a multiple birth were clustered within families. For illustration purposes the clustering in DINO was removed by excluding second and subsequent born infants from a multiple birth from the analysis dataset; such an approach would not be recommended in practice. Since ignoring clustering can lead to biased standard errors for parameter estimates (149), it is important that clustering is accounted for in the analysis. When the analysis model allows for clustering, for example in a mixed effects model or using generalised estimating equations, the imputation model should also account for the clustering. Excluding the case of longitudinal data, where missing data can be imputed by treating the different measurements over time as different variables in the dataset (i.e. data in wide format), accounting for clustering in the imputation model can be a challenging

process. Potential approaches include treating cluster as a fixed effect within the imputation model, imputing separately by cluster, or fitting a multilevel imputation model (5, 15). Although it is expected that the main findings of this thesis apply to these more complex types of imputation models, this is a topic for future research.

Another limitation of this thesis is that the missing data problem in the DINO case study was not severe. As described in chapters 6 and 7 of the thesis, treatment effect estimates for fat free mass at 7 years corrected age were similar regardless of how MI was implemented, whether performed overall, separately by randomised group, in combination with IPW, or restricted to children who provided outcome data during the extended follow-up phase of the trial. Further, despite the inclusion of auxiliary variables in the imputation model, treatment effect estimates did not substantially differ between MI and a simple adjusted CCA. Although in one sense it was reassuring to note that treatment effect estimates were consistent across approaches, it would be informative to consider case studies where results are more sensitive to the choice of approach for handling missing outcome data.

8.3. Concluding remarks

As highlighted at the beginning of this thesis, MI is a flexible and increasingly popular statistical approach for handling missing data. Despite a growing evidence base for its use, implementation in practical settings remains challenging, and in many cases there is no consensus in the literature to guide decisions around how to best generate imputed datasets for analysis. This thesis has focused on knowledge gaps in the application of MI for handling missing outcome data, which is a common problem in medical research. In particular, this thesis has explored the use of MI for handling missing outcome data in the presence of auxiliary variables for the outcome, when estimating relative risks, and in RCTs and extended follow-up studies based on RCTs. The research has demonstrated the benefits of retaining imputed outcomes for analysis, the

shortcomings of standard model-based methods of MI for estimating the relative risk, and the limited utility of MI in some RCT settings. In addition, this thesis has offered guidance on how imputation models should be specified in the context of RCTs and extended follow-up studies. Findings and recommendations from this work will enable researchers to make more informed decisions about the appropriate implementation of MI for handling missing outcome data in applied settings.

9. References

1. National Research Council, Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The prevention and treatment of missing data in clinical trials. Washington (DC): National Academies Press (US); 2010.
2. Rubin D. Inference and missing data. *Biometrika*. 1976;63(3):581-92.
3. Rubin D, editor Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association; 1978.
4. Rubin D. Multiple imputation for nonresponse in surveys. New York: Wiley & Sons; 1987.
5. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011;30(4):377-99.
6. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. 2001;6(4):330-51.
7. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*. 2015;15:30.
8. von Hippel PT. Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*. 2007;37(1):83-117.
9. Raghunathan T, Lepkowski J, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 2001;27(1):85-95.
10. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007;16(3):219-42.

11. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
12. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*. 2004;1(4):368-76.
13. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Statistical Methods in Medical Research*. 2014;23(5):440-59.
14. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods*. 2002;7(2):147-77.
15. Graham JW. Missing data analysis: making it work in the real world. *Annual Review of Psychology*. 2009;60:549-76.
16. Drye LT, Casper AS, Sternberg AL, Holbrook JT, Jenkins G, Meinert CL. The transitioning from trials to extended follow-up studies. *Clinical Trials*. 2014;11(6):635-47.
17. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*. 2009;338:b2393.
18. Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association*. 1992;87(420):1227-37.
19. Carpenter J, Kenward M. Missing data in randomised controlled trials - a practical guide. Birmingham: National Institute for Health Research; 2007.
20. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical Trials*. 2012;9(4):396-407.
21. Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research*. 2013;25(4):1471-89.
22. Little R, Rubin D. *Statistical analysis with missing data*, 2nd edition. New York: Wiley; 2002.

23. Graham JW, Donaldson SI. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*. 1993;78(1):119-28.
24. Groenwold RH, Donders AR, Roes KC, Harrell FE, Jr., Moons KG. Dealing with missing outcome data in randomized trials and observational studies. *American Journal of Epidemiology*. 2012;175(3):210-7.
25. Little RJ, Rubin D. *Statistical analysis with missing data*. New Jersey: John Wiley & Sons; 1987.
26. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. 2013;22(3):278-95.
27. Schafer JL. Multiple imputation: a primer. *Statistical methods in Medical Research*. 1999;8(1):3-15.
28. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*. 2007;16(3):199-218.
29. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*. 2007;26(16):3057-77.
30. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*. 2015;24(4):462-87.
31. Morris TP, White IR, Royston P, Seaman SR, Wood AM. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine*. 2014;33(1):88-104.
32. von Hippel PT. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*. 2009;39(1):265-91.
33. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*. 2012;12:46.
34. White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in Medicine*. 2009;28(15):1982-98.

35. Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. *Statistics in Medicine*. 2015;34(25):3298-317.
36. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. 1998;33(4):545-71.
37. Romaniuk H, Patton GC, Carlin JB. Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods. *American Journal of Epidemiology*. 2014;180(9):920-32.
38. von Hippel PT. Should a normal imputation model be modified to impute skewed variables? *Sociological Methods and Research*. 2013;42(1):105-38.
39. Lee KJ, Carlin JB. Multiple imputation in the presence of non-normal data. *Statistics in Medicine*. 2017;36(4):606-17.
40. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*. 2007;26(6):1368-82.
41. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*. 2010;171(5):624-32.
42. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*. 2011;20(1):40-9.
43. van Buuren S, Brand J, Groothuis-Oudshoorn C, Rubin D. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*. 2006;76(12):1049-64.
44. Allison P. *Missing data*. Thousand Oaks, California: Sage Publications Inc; 2002.
45. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*. 2007;8(3):206-13.

46. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. 1999;18(6):681-94.
47. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*. 2012;367(14):1355-60.
48. Graham JW. *Missing data: analysis and design*: Springer New York; 2012.
49. Mustillo S. The effects of auxiliary variables on coefficient bias and efficiency in multiple imputation. *Sociological Methods & Research*. 2012;41(2):335-61.
50. Howard WJ, Rhemtulla M, Little TD. Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*. 2015;50(3):285-99.
51. Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology*. 2012;12:184.
52. Yoo JE. The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educational and Psychological Measurement*. 2009;69:929-47.
53. Doidge JC. Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at random. *Statistical Methods in Medical Research*. 2016.
54. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*. 1987;125(5):761-8.
55. Lee J. Odds ratio or relative risk for cross-sectional data? *International Journal of Epidemiology*. 1994;23(1):201-3.
56. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology*. 1994;47(8):881-9.

57. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*. 2003;157(10):940-3.
58. Cummings P. The relative merits of risk ratios and odds ratios. *Archives of Pediatrics & Adolescent Medicine*. 2009;163(5):438-45.
59. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*. 2004;13(4):309-23.
60. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *American Journal of Epidemiology*. 1986;123(1):174-84.
61. Skov T, Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. *International Journal of Epidemiology*. 1998;27(1):91-5.
62. Zou G. A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*. 2004;159(7):702-6.
63. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology*. 2003;3:21.
64. Muller CJ, MacLehose RF. Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *International Journal of Epidemiology*. 2014;43(3):962-70.
65. ICH E9 Expert Working Group. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Statistics in Medicine*. 1999;18(15):1905-42.
66. Permutt T. A taxonomy of estimands for regulatory clinical trials with discontinuations. *Statistics in Medicine*. 2016;35(17):2865-75.
67. Mallinckrodt C, Molenberghs G, Rathmann S. Choosing estimands in clinical trials with missing data. *Pharmaceutical Statistics*. 2017;16(1):29-36.
68. Little R, Kang S. Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. *Statistics in Medicine*. 2015;34(16):2381-90.

69. Heritier SR, Gebski VJ, Keech AC. Inclusion of patients in clinical trial analysis: the intention-to-treat principle. *Medical Journal of Australia*. 2003;179(8):438-40.
70. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal*. 2010;340:c869.
71. Committee for Proprietary Medicinal Products. Guideline on missing data in confirmatory clinical trials. 2009 EMA/CPMP/EWP/1776/99 Rev. 1.
72. Committee for Proprietary Medicinal Products. Points to consider on missing data. 2001 CPMP/EWP/1776/99.
73. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. *Open Medicine*. 2009;3(2):e51-3.
74. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale Journal of Biology and Medicine*. 2013;86(3):343-58.
75. Leon AC, Demirtas H, Hedeker D. Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial. *Clinical Trials*. 2007;4(5):540-7.
76. Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*. 2013;12(6):337-47.
77. Groenwold RH, Moons KG, Vandenbroucke JP. Randomized trials with missing outcome data: how to analyze and what to report. *Canadian Medical Association journal*. 2014;186(15):1153-7.
78. Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Statistics in Medicine*. 2002;21(8):1043-66.
79. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38(4):963-74.
80. Wang C, Hall CB. Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine*. 2010;29(6):671-9.

81. Chen Q, Ibrahim JG. A note on the relationships between multiple imputation, maximum likelihood and fully Bayesian methods for missing responses in linear regression models. *Statistics and its Interface*. 2014;6(3):315-24.
82. Allison PD. Handling missing data by maximum likelihood. *SAS Global Forum* 2012.
83. Cuzick J. Long-term follow-up in cancer prevention trials (It ain't over 'til it's over). *Cancer Prevention Research*. 2010;3(6):689-91.
84. Goodman PJ, Hartline JA, Tangen CM, Crowley JJ, Minasian LM, Klein EA, et al. Moving a randomized clinical trial into an observational cohort. *Clinical Trials*. 2013;10(1):131-42.
85. Baer JS, Kivlahan DR, Blume AW, McKnight P, Marlatt GA. Brief intervention for heavy-drinking college students: 4-year follow-up and natural history. *American Journal of Public Health*. 2001;91(8):1310-6.
86. Fergusson DM, Boden JM, Horwood LJ. Nine-year follow-up of a home-visitation program: a randomized trial. *Pediatrics*. 2013;131(2):297-303.
87. Makrides M, Gould JF, Gawlik NR, Yelland LN, Smithers LG, Anderson PJ, et al. Four-year follow-up of children born to women in a randomized trial of prenatal DHA supplementation. *Journal of the American Medical Association*. 2014;311(17):1802-4.
88. Martin RM, Patel R, Kramer MS, Guthrie L, Vilchuck K, Bogdanovich N, et al. Effects of promoting longer-term and exclusive breastfeeding on adiposity and insulin-like growth factor-I at age 11.5 years: a randomized trial. *Journal of the American Medical Association*. 2013;309(10):1005-13.
89. Lindstrom J, Peltonen M, Eriksson JG, Ilanne-Parikka P, Aunola S, Keinanen-Kiukaanniemi S, et al. Improved lifestyle and decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish Diabetes Prevention Study (DPS). *Diabetologia*. 2013;56(2):284-93.
90. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine*. 2006;25(24):4279-92.
91. Makrides M, Gibson RA, McPhee AJ, Collins CT, Davis PG, Doyle LW, et al. Neurodevelopmental outcomes of preterm infants fed high-dose

- docosahexaenoic acid: a randomized controlled trial. *Journal of the American Medical Association*. 2009;301(2):175-82.
92. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*. 2010;29(28):2920-31.
 93. Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. *Emerging Themes in Epidemiology*. 2012;9(1):3.
 94. Brunner EJ, Shipley MJ, Witte DR, Singh-Manoux A, Britton AR, Tabak AG, et al. Arterial stiffness, physical function, and functional limitation: the Whitehall II Study. *Hypertension*. 2011;57(5):1003-9.
 95. Horn K, Dino G, Branstetter SA, Zhang J, Noerachmanto N, Jarrett T, et al. Effects of physical activity on teen smoking cessation. *Pediatrics*. 2011;128(4):e801-11.
 96. Bot M, Pouwer F, Zuidersma M, van Melle JP, de Jonge P. Association of coexisting diabetes and depression with mortality after myocardial infarction. *Diabetes Care*. 2012;35(3):503-9.
 97. Zuidersma M, Conradi HJ, van Melle JP, Ormel J, de Jonge P. Self-reported depressive symptoms, diagnosed clinical depression and cardiac morbidity and mortality after myocardial infarction. *International Journal of Cardiology*. 2013;167(6):2775-80.
 98. Hiyoshi A, Fukuda Y, Shipley MJ, Brunner EJ. Inequalities in self-rated health in Japan 1986-2007 according to household income and a novel occupational classification: national sampling survey series. *Journal of Epidemiology and Community Health*. 2013;67(11):960-5.
 99. Hafkamp-de Groen E, Lingsma HF, Caudri D, Levie D, Wijga A, Koppelman GH, et al. Predicting asthma in preschool children with asthma-like symptoms: validating and updating the PIAMA risk score. *Journal of Allergy and Clinical Immunology*. 2013;132(6):1303-10.
 100. Lee H, Andrew M, Gebremariam A, Lumeng JC, Lee JM. Longitudinal associations between poverty and obesity from birth through adolescence. *American Journal of Public Health*. 2014;104(5):e70-6.

101. Hatzenbuehler ML, Bellatorre A, Muennig P. Anti-gay prejudice and all-cause mortality among heterosexuals in the United States. *American Journal of Public Health*. 2014;104(2):332-7.
102. Carpenter JR, Kenward MG. *Multiple imputation and its application*. Chichester, UK: Wiley & Sons; 2013.
103. Kontopantelis E, White IR, Sperrin M, Buchan I. Outcome-sensitive multiple imputation: a simulation study. *BMC Medical Research Methodology*. 2017;17(1):2.
104. Sullivan TR, Salter AB, Ryan P, Lee KJ. Bias and precision of the "Multiple Imputation, Then Deletion" method for dealing with missing outcome data. *American Journal of Epidemiology*. 2015;182(6):528-34.
105. Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. *American Journal of Epidemiology*. 2011;174(8):984-92.
106. Miettinen OS, Cook EF. Confounding: essence and detection. *American Journal of Epidemiology*. 1981;114(4):593-603.
107. Little RJ, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Neaton JD, et al. The design and conduct of clinical trials to limit missing data. *Statistics in Medicine*. 2012;31(28):3433-43.
108. Fleming TR. Addressing missing data in clinical trials. *Annals of Internal Medicine*. 2011;154(2):113-7.
109. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*. 2014;14(1):118.
110. Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004;5(3):445-64.
111. Mackinnon A. The use and reporting of multiple imputation in medical research - a review. *Journal of Internal Medicine*. 2010;268(6):586-93.
112. Carpenter JR, Roger JH, Kenward MG. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and

- inference via multiple imputation. *Journal of Biopharmaceutical Statistics*. 2013;23(6):1352-71.
113. Alshurafa M, Briel M, Akl EA, Haines T, Moayyedi P, Gentles SJ, et al. Inconsistent definitions for intention-to-treat in relation to missing outcome data: systematic review of the methods literature. *PLoS One*. 2012;7(11):e49163.
 114. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*. 2004;57(5):454-60.
 115. Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15:139.
 116. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*. 2011;173(7):731-8.
 117. Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*. 2005;2(5):379-86.
 118. Mallinckrodt CH, Clark SW, Carroll RJ, Molenbergh G. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics*. 2003;13(2):179-90.
 119. Lu K, Mehrotra DV. Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Statistics in Medicine*. 2010;29(4):474-88.
 120. Littell RC, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*. 2000;19(13):1793-819.
 121. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53(3):983-97.

122. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*. 2005;24(7):993-1007.
123. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*. 2006;59(10):1087-91.
124. Knol MJ, Janssen KJ, Donders AR, Egberts AC, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology*. 2010;63(7):728-36.
125. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Canadian Medical Association journal*. 2012;184(11):1265-9.
126. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology*. 1991;44(1):77-81.
127. Collins CT, Makrides M, Gibson RA, McPhee AJ, Davis PG, Doyle LW, et al. Pre- and post-term growth in pre-term infants supplemented with higher-dose DHA: a randomised controlled trial. *British Journal of Nutrition*. 2011;105(11):1635-43.
128. Collins CT, Gibson RA, Anderson PJ, McPhee AJ, Sullivan TR, Gould JF, et al. Neurodevelopmental outcomes at 7 years' corrected age in preterm infants who were fed high-dose docosahexaenoic acid to term equivalent: a follow-up of a randomised controlled trial. *BMJ Open*. 2015;5(3):e007314.
129. Mazza GL, Enders CK, Ruehlman LS. Addressing item-level missing data: a comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research*. 2015;50(5):504-19.
130. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44(4):1049-60.

131. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*. 2009;24(3):127-35.
132. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*. 2014;15:237.
133. Fleiss JL, Levin B, Paik MC. *Statistical methods for rates and proportions*. 3rd ed. New Jersey: John Wiley & Sons; 2003.
134. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*. 2014;11(5):590-600.
135. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*. 2016;17:72.
136. Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. Available from www.cochrane-handbook.org; The Cochrane Collaboration; 2011.
137. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37-46.
138. Paik MC. The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*. 1997;92(440):1320-9.
139. Finfer S, Chittock D, Li Y, Foster D, Dhingra V, Bellomo R, et al. Intensive versus conventional glucose control in critically ill patients with traumatic brain injury: long-term follow-up of a subgroup of patients from the NICE-SUGAR study. *Intensive Care Medicine*. 2015;41(6):1037-47.
140. Shih WJ. Comments on the three papers by the FDA/CDER research team on the regulatory perspective of the missing data problem. *Statistics in Medicine*. 2016;35(17):2880-6.
141. McKelvey L, Schiffman RF, Brophy-Herb HE, Bocknek EL, Fitzgerald HE, Reischl TM, et al. Examining long-term effects of an infant mental health

- home-based early head start program on family strengths and resilience. *Infant Mental Health Journal*. 2015;36(4):353-65.
142. Desai AA, Alemayehu H, Holcomb GW, 3rd, St Peter SD. Minimal vs. maximal esophageal dissection and mobilization during laparoscopic fundoplication: long-term follow-up from a prospective, randomized trial. *Journal of Pediatric Surgery*. 2015;50(1):111-4.
143. Molyneux AJ, Birks J, Clarke A, Sneade M, Kerr RS. The durability of endovascular coiling versus neurosurgical clipping of ruptured cerebral aneurysms: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT). *Lancet*. 2015;385(9969):691-7.
144. McNamara HC, Wood R, Chalmers J, Marlow N, Norrie J, MacLennan G, et al. STOPPIT Baby Follow-up Study: the effect of prophylactic progesterone in twin pregnancy on childhood outcome. *PLoS One*. 2015;10(4):e0122341.
145. Andersen LL, Ottesen B, Alling Moller LM, Gluud C, Tabor A, Zobbe V, et al. Subtotal versus total abdominal hysterectomy: randomized clinical trial with 14-year questionnaire follow-up. *American Journal of Obstetrics and Gynecology*. 2015;212(6):758.e1-.e54.
146. Secher RG, Hjorthoj CR, Austin SF, Thorup A, Jeppesen P, Mors O, et al. Ten-year follow-up of the OPUS specialized early intervention trial for patients with a first episode of psychosis. *Schizophrenia Bulletin*. 2015;41(3):617-26.
147. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58(1):21-9.
148. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics*. 2012;68(1):129-37.
149. Cannon MJ, Warner L, Taddei JA, Kleinbaum DG. What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. *Statistics in Medicine*. 2001;20(9-10):1461-7.