Exploring strategies *for* incorporating population-level external information *in* multiple imputation *of* missing data

Tra My Pham

Thesis submitted for the degree of Doctor of Philosophy

University College London

I, Tra My Pham, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

This PhD is supported by awards to establish the Farr Institute of Health Informatics Research, London, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, NIHR, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1).

Abstract

Multiple imputation (MI) is increasingly used for handling missing data in medical research. The standard implementation of MI assumes that data are missing at random (MAR). However, under missing not at random (MNAR) mechanisms, standard MI might not be satisfactory.

When there are external data sources providing population-level information about the incomplete variables, it is desirable to utilise such information in MI. This thesis aims to explore how knowledge about the incomplete covariate's population marginal distribution from an external dataset can be used to improve standard MI under MNAR mechanisms. Two univariate MI methods are proposed for an incomplete binary/categorical covariate to anchor inference to the population: *weighted MI* and *calibrated-δ adjustment MI*.

Chapter 3 demonstrates how, in weighted MI, the incomplete covariate's population distribution can be incorporated as probability weights in the imputation process to closely match the post-imputation distribution to the population level. Results from analytic and simulation studies of a 2×2 contingency table show that weighted MI can produce more accurate inferences under two general MNAR mechanisms. Weighted MI is also integrated into the multivariate imputation by chained equations (MICE) algorithm for imputing several incomplete covariates, accounting for their population marginal distributions from external data.

Chapter 4 develops and evaluates calibrated-δ adjustment MI, which incorporates the incomplete covariate's population distribution as a δ adjustment in the imputation model's intercept. In a 2×2 contingency table, it is shown analytically and via simulation that appropriately adjusting the imputation model's intercept fully corrects bias when the incomplete covariate is MNAR dependent on its values and the (complete) outcome. An adaptation of the method in the MICE algorithm for multivariate imputation is also explored.

Chapter 5 investigates another univariate missing data setting, with a continuous outcome. Under the above MNAR mechanism, the presence of a second sensitivity parameter for the covariate–outcome association in the imputation model is introduced, rendering the calibrated-δ intercept adjustment insufficient. The sensitivity analysis then involves eliciting values of the second sensitivity parameter and deriving the calibrated-δ adjustment in the intercept.

Chapter 6 presents two case studies using electronic health records to illustrate the application of the proposed population-calibrated MI methods.

Outline

1	Introduction
2	Missing data and multiple imputation
3	Weighted multiple imputation of a binary covariate when the outcome variable is binary
4	Calibrated-δ adjustment multiple imputation of a binary covariate when the outcome variable is binary
5	Population-calibrated multiple imputation of a binary covariate when the outcome variable is continuous
6	Case studies using UK primary care electronic health records
7	Discussion
§	Appendices

Acknowledgements

My PhD journey has been such an enjoyable learning experience, thanks to many people.

Irene Petersen, my principal supervisor, has given me the freedom to pursue my own statistical interest, while helping me keep the bigger picture. The completion of this thesis would not have been achievable without Irene's thoughtful guidance, encouragement, and constructive feedback.

James Carpenter, my subsidiary supervisor, has been exceptional with his guidance and input. His insightful questions and stimulating discussions have been the main source for me to work out the ideas presented in this thesis. It has been an inspiring experience working with James.

Tim Morris, my subsidiary supervisor, has been continuously engaged in my work and encouraging throughout. Tim has been tremendously generous with his time, and provided much help with several issues on programming and conducting simulation studies.

Angela Wood, *Ian White*, and *Kate Tilling* have offered valuable ideas and perspectives, and I thank them for their interest in my work.

Colleagues at the *Department of Primary Care & Population Health, UCL* have created a warm and welcoming research environment. Special thanks go to *Manuj Sharma* for sharing his insights and expertise, particularly his work on type 2 diabetes diagnoses in primary care. I would also like to thank *Ann Liljas, Yifeng Liu, Sonia Coton, April Slee*, and *Hilary Davies* for their help with proofreading my thesis.

My *family* and *friends* have been incredible pillars of strength and comfort. I am forever grateful to my *Dad*, who has given me the opportunities to be where I am today. I would like to acknowledge the unwavering support of *Valentina, Kavya, Theodoros, Lorenzo*, and *Rafael*; the laughs we shared have kept me sane. I am blessed with the love and constant encouragement from my partner and best friend *Linh*.

Finally, this thesis is dedicated to my *Mum*, to whom I owe everything.

Contents

List of figures

List of tables

Abbreviations

ACU	Acceptable computer usage
AHD	Additional health data
AMR	Acceptable mortality recording
CI	Confidence interval
CPRD	Clinical Practice Research Datalink
CRA	Complete record analysis
FCS	Fully conditional specification
FMI	Fraction of missing information
GP	General practitioner
GPRD	General Practice Research Database
IMR	Inverse Mills ratio
JAV	Just another variable
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MCSE	Monte Carlo standard error
MI	Multiple imputation
MICE	Multivariate imputation by chained equations
MNAR	Missing not at random
NHS	National Health Service
ONS	Office for National Statistics
OR	Odds ratio
QOF	Quality and Outcomes Framework
RE	Relative efficiency
RRR	Relative risk ratio
RVI	Relative increase in variance
SD	Standard deviation
SE	Standard error
SMC	Substantive model compatible
THIN	The Health Improvement Network
UK	United Kingdom

Notations

α	parameter(s) of the selection model
β	parameter(s) of the analysis model
δ	sensitivity parameter in MI under MNAR
ε	residual error
θ	parameter(s) of the imputation model
μ	mean
ν	degree of freedom
ρ	correlation
σ	standard deviation
$\phi\,()$	probability density function of the standard normal distribution
$\Phi\,()$	cumulative distribution function of the standard normal distribution
a	a real-number value
b	cell-wise bias in the analytic study of a 2×2 contingency table
B	between-imputation variance
c	denotes quantities obtained in conditional weighted MI
corr ()	correlation coefficient
C	coverage of 95% confidence intervals
d	derivative
e	a random draw from a χ^2 distribution
ex	denotes quantities obtained in the external dataset
exp()	exponential function
expit()	inverse of the logit function, expit $(a) = \frac{1}{1+\exp(-a)}$
E ()	expectation
$f\,()$	function
full	denotes quantities obtained in the full data
h	denotes covariates in the Heckman selection model
H	hypothesis
i	indexes individuals in the dataset
$I\,()$	indicator function
j	indexes variables in the dataset
k	indexes categories of the K-level categorical variable
l	indexes categories of the L-level categorical variable
linpred	linear predictor
ln()	natural logarithm
$L\,()$	likelihood function
m	indexes imputations in MI

m	denotes quantities obtained in marginal weighted MI
mid	denotes the midpoint of an interval
mis	denotes quantities in the missing data
M	number of imputations in MI
MVN $(\,)$	multivariate normal distribution
n	number of individuals/observations in the dataset
N$(\,)$	normal distribution
obs	denotes quantities in the observed data
p	number of variables; or probability
pop	denotes population-level quantities
pred	denotes predicted quantities in the completed data
ps	post-stratification
$p\,(\,)$	probability; or distribution (section 2.4.2)
q	number of incomplete variables
r	response indicator of an incomplete variable(s)
R^2	coefficient of determination
req	denotes quantities required to be imputed in the missing data
s	denotes quantities obtained in standard MI
s	indexes simulation repetitions
S	number of simulations
t	t-distribution
(t)	indexes iterations in an algorithm
T	number of iterations
u	a random draw from the standard normal distribution
Var$(\,)$	variance
w	probability weight
W	within-imputation variance
x	covariate(s) in the analysis model
y	outcome variable in the analysis model
z	random variable(s)

$\hat{}$	denotes an estimate of the corresponding parameter underneath
$\dot{}$	denotes a random draw
$*$	denotes the combined parameter
$'$	denotes the transpose of a matrix
$+$	denotes the sum over rows or columns of the contingency table
$\bar{}$	denotes the mean of the corresponding quantity underneath

1

Introduction

1.1 BACKGROUND AND MOTIVATION

Primary care databases of electronic health records containing routinely collected clinical information about patients in primary care have been recognised as rich data sources for health research. In the United Kingdom (UK), there are several large primary care databases which typically hold data collected from several hundred general practices across the UK since the late 1980s. These databases offer many opportunities for research on populations that are otherwise difficult to recruit in clinical trials or cohort studies, such as individuals with severe mental illness [1, 2], pregnant women [3–5], children [6], and the elderly [7, 8].

One example of primary care electronic health records is the The Health Improvement Network (THIN) database. THIN contains longitudinal electronic health records of more than 12 million patients registered with over 600 general practices in the UK. Data are collected from the point of practice registration to the time the patients leave or die. Information captured in THIN includes medical diagnoses, symptoms, prescribed medication, health indicators, and lifestyle factors recorded through patient consultations with the general practitioners or healthcare professionals in primary care.

The recording of variables in primary care databases generally reflects how patient information is collected in the primary care setting. During the first year of registration with the general practices, most patients have a record of common information, including past and current medical history as well as measurements of some health indicators and lifestyle factors such as height, weight, blood pressure, smoking status, and alcohol consumption. Data are not systematically recorded or updated thereafter, unless they are directly relevant for patient management and care. As a result, data in variables required for research purposes are often incomplete or missing, which can obstruct their use in primary care research.

Ethnicity is associated with disparities in disease prevalences and healthcare utilisation, and is an important factor to be considered in many epidemiological studies [9–11]. Within primary

care, the facility to record ethnicity has been introduced since 1991, and therefore ethnicity data are available in several UK primary care databases including THIN [12]. The completeness of ethnicity recording is ideal when primary care databases are used to investigate ethnic differences in disease epidemiology, healthcare utilisation, or outcomes. Unfortunately, there is a large amount of missing data in ethnicity in such databases [13], which quite severely limits the use of ethnicity information recorded in primary care in research. Although ethnicity is associated with a number of important health conditions, such as type 2 diabetes [11], cardiovascular diseases [10], and severe mental illness [9], many studies using primary care databases are often reluctant to include ethnicity in the analysis, primarily due to the low level of recording [1, 2, 14].

The recording of ethnicity information in primary care has a direct impact on the level of missing data in ethnicity in primary care databases, which is of relevant concern for research using ethnicity data in such databases. However, only a few studies have investigated the completeness of ethnicity information in primary care. Kumarapeli et al. [15] analysed the recording of ethnicity in 16 general practices before and after an intervention which targeted at improving the completeness of ethnicity data. The authors reported a poor baseline recording of ethnicity data, with less than 1% of the practice population having ethnicity codes recorded prior to the intervention, and that the median level of ethnicity recording increased to 47% after the intervention. Mathur et al. [13] examined the recording of ethnicity information in the Clinical Practice Research Datalink (CPRD) primary care database and found that less than 30% of all individuals in the database (1990–2012) had a record of ethnicity. Ethnicity recording was included in the 2006/7 revision of the Quality and Outcomes Framework (QOF), which provided general practices with a financial incentive to record ethnicity in all new patient registrations. Following the financial incentivisation of ethnicity recording under QOF, the completeness of ethnicity data for newly registered patients was improved immensely. Indeed, according to Mathur et al. [13], the percentage of individuals with a record of ethnicity increased from approximately 20–30% for those first registered prior to the QOF financial incentivisation in 2006/7, to 70–80% for those who registered after 2006/7. Ethnicity recording was later removed from QOF in the 2011/12 update, and since then no studies have investigated whether there was a subsequent drop in ethnicity recording. As Mathur et al. [13]'s study analysed CPRD data up to 2011, they were not able to assess the impact of the removal of ethnicity recording from the QOF scheme in the 2011/12 financial year. However, the recording of ethnicity is anecdotally expected to decline following the 2011/12 QOF update.

Several simple, or 'ad-hoc', methods have been employed for the analysis of missing data in research using clinical databases. Such methods include complete record analysis (CRA, in which only individuals with complete information on all variables considered are included in the analysis), excluding variables with missing data from the analysis, and single-value imputation techniques. The issue of bias and potentially misleading conclusions associated with these methods is well-known [16–18]. Within the context of incomplete ethnicity data, if the characteristics of individuals whose ethnicity is recorded are systematically different from those whose ethnicity is missing, then a CRA ignoring individuals with missing ethnicity information can potentially lead to biased results. In addition, given the large proportions of individuals with incomplete ethnicity information in primary care databases, a CRA may lead to a substantial reduction in

sample size and power. The omission of ethnicity from the analysis model may introduce bias if ethnicity is known to be associated with the health outcome of interest. Alternatively, some previous studies made the assumption that all individuals with missing ethnicity data belonged to the White ethnic group (i.e. only White individuals ever had missing ethnicity information) [19]. Replacing all missing values in ethnicity with the White ethnic group might lead to some non-White individuals being misclassified, which in turn might dilute any effect of the different ethnic groups on the outcome of interest. In addition, this approach might misrepresent the ethnic make-up of the sample, with one possible consequence being that primary care databases appear less 'representative' of the population than they actually are.

Multiple imputation (MI) [20] is increasingly regarded as the standard approach for dealing with missing data in medical research [18]. Unlike the complete record analysis, MI utilises information from individuals with incomplete data. Thus, MI can produce unbiased and statistically more powerful analyses compared to other simple methods [18]. In MI, each missing value is replaced with several plausible values generated from an imputation model, conditional on the observed data. This procedure creates a number of completed datasets to account for the uncertainty introduced by missing data. The desired analysis is then performed in each of these completed datasets. Finally, the resulting parameter estimates and standard errors are combined into a single set of results using Rubin's rules [20, 21], taking into account the variation within and between the datasets.

In practice, MI is commonly implemented under the assumption of data being missing completely at random (MCAR, when missingness does not depend on either observed or unobserved information), or missing at random (MAR, when missingness does not depend on unobserved information, conditional on observed information). However, it is possible that data are missing not at random (MNAR, when missingness depends on unobserved information, even after conditioning on observed information). In primary care databases, missing data in ethnicity may depend on unobserved information, such as the unrecorded ethnic groups, or other factors affecting the recording of ethnicity that are not accessible in the databases [13]. This implies a potential underlying MNAR mechanism for ethnicity, and as a result standard MI assuming MAR might not cope. In particular, standard MI might fail to yield a plausible estimate of the marginal distribution of ethnicity.

For an incomplete variable in a given dataset, its corresponding population-level marginal distribution might also be available in an external data source. As an example, for ethnicity data in large UK primary care databases, the corresponding distribution of ethnicity in the population is obtainable from the UK census statistics. It is therefore natural to incorporate this external information in the imputation process, assuming that the study sample should be representative of the external population data in terms of the incomplete variable. If done appropriately, the inclusion of such knowledge about the incomplete variable can potentially improve on standard MI for missing data generated by general MNAR mechanisms.

From a methodological point of view, statistical research has proposed methods for the analysis of incomplete data under the assumption of data being MNAR. However, these methods are often not calibrated (with the exception of Carpenter et al.'s reference based sensitivity analysis [22]). This thesis outlines methods for calibrating MI inferences to the population level under

general MNAR assumptions for missing data in the above applied setting. The overall aim of this thesis is to systematically investigate how an external data source containing population-level information about the incomplete variable can be appropriately utilised in the imputation process to improve standard MI when missing data are suspected to be MNAR, as detailed below.

1.2 AIMS AND OBJECTIVES

This thesis aims to explore the use of available external data sources containing the population-level marginal distribution of the incomplete variable in improving standard MI under general MNAR mechanisms. Motivated by the issue of incomplete ethnicity information in primary care databases, this thesis focuses on MI methods for accommodating missing data in incomplete binary/categorical variables which are included as covariates in the analysis model.

Two univariate population-calibrated MI methods which incorporate knowledge about the incomplete covariate's population distribution in MI are proposed to calibrate inference to the population: *weighted multiple imputation* and *calibrated-δ adjustment multiple imputation*. In weighted MI, the incomplete covariate's population distribution is used to calculate probability weights, which are then used in the imputation process to closely match the post-imputation distribution to the population level. Alternatively, the calibrated-δ adjustment MI method incorporates the incomplete covariate's population distribution as a δ adjustment in the intercept of the imputation model for the covariate.

The univariate population-calibrated MI methods can also be integrated into the multivariate imputation by chained equations (MICE) algorithm [23] to impute missing values in several incomplete covariates, accounting for their marginal distributions in population data. MICE fills in missing values in the incomplete variables iteratively by using chained equations, a sequence of univariate imputation models which are specified for each of the incomplete variables conditional on all the other variables.

More specifically, the objectives of this thesis are as follows.

1. To develop and evaluate weighted MI of an incomplete binary covariate, and to explore the inclusion of univariate weighted MI in the MICE algorithm for imputing missing values in several incomplete binary covariates, when the outcome variable is binary;
2. To develop and evaluate calibrated-δ adjustment MI of an incomplete binary covariate, and to explore the inclusion of univariate calibrated-δ adjustment MI in the MICE algorithm for imputing missing values in several incomplete binary covariates, when the outcome variable is binary;
3. To evaluate calibrated-δ adjustment MI and weighted MI of an incomplete binary covariate when the outcome variable is continuous;
4. To implement calibrated-δ adjustment MI and weighted MI for handling missing data in ethnicity in case studies using UK primary care electronic health records.

The first three objectives are achieved by performing a series of analytic and simulation studies, using increasingly complex missingness mechanisms for the incomplete covariate(s). Throughout these studies, the population-calibrated MI methods are also compared to standard MI and complete record analysis (CRA). The last objective is achieved by conducting two ethnicity-focused case studies using THIN data. In these case studies, the population-calibrated

MI methods are compared to standard MI and other simple approaches to missing data.

The next section provides an outline of the subsequent chapters in this thesis.

1.3 OUTLINE OF SUBSEQUENT CHAPTERS

Chapter 2 presents an overview of the issues raised by missing data in medical research and the available methods for handling missing data. This chapter outlines the different missingness patterns and mechanisms which are the key concepts in the analysis of incomplete data. Simple methods for handling missing data are reviewed, before multiple imputation and various aspects of multiple imputation analysis are introduced.

Chapter 3 proposes and evaluates the *weighted multiple imputation* method for utilising external information about the incomplete variable's population distribution in MI, in order to calibrate inference to the population. This chapter describes the procedure of the univariate weighted MI method, as well as the derivation of the marginal and conditional weights from the incomplete variable's population distribution. These weights are used in weighted MI to recover the correct incomplete variable's distribution after imputation. Weighted MI is evaluated and compared to standard MI and CRA in analytic and simulation studies of a 2×2 contingency table, with a complete binary outcome variable and an incomplete binary covariate. The investigation is then extended to a multivariate missing data setting. Univariate weighted MI is integrated into the MICE algorithm, and this integration is evaluated and compared to standard MICE and CRA in multivariate simulation studies. These studies feature a three-way contingency table with a fully observed binary outcome variable and two partially observed binary covariates, and different missingness mechanisms for the covariates are considered.

Chapter 4 proposes and evaluates the *calibrated-δ adjustment multiple imputation* method as an alternative approach to weighting in MI when the population-level marginal distribution of the incomplete variable is available externally. This method is motivated by van Buuren et al.'s δ adjustment (offset) MI method [23]. In calibrated-δ adjustment MI, the incomplete variable's population distribution is used (together with its observed-data distribution and association with other fully observed variables) to calculate an adjustment in the imputation model's intercept. The univariate missing data setting of a 2×2 contingency table discussed in chapter 3 is revisited. Calibrated-δ adjustment MI is evaluated and compared to weighted MI, standard MI, and CRA analytically and via simulation. Univariate calibrated-δ adjustment MI is also adapted for use in the MICE algorithm. This adaptation is further evaluated and compared to the integration of univariate weighted MI in MICE, standard MICE, and CRA in multivariate simulation studies of the same set-up as outlined in chapter 3.

Chapter 5 investigates a univariate missing data setting where the incomplete covariate is binary as before, but the complete outcome variable is continuous. The population-calibrated MI methods are evaluated and compared to standard MI and CRA in a univariate simulation study. A proof-of-concept example based on the ideas of the Heckman model [24] is also conducted to provide theoretical support for the empirical results of the univariate simulation study. The last part of chapter 5 brings together these empirical and theoretical findings to evaluate and compare population-calibrated MI, standard MI, and CRA in further simulations.

Chapter 6 illustrates the application of calibrated-δ adjustment MI and weighted MI in real-

life settings, using UK primary care electronic health records. In two case studies conducted using THIN data, these population-calibrated MI methods are implemented for handling missing data in ethnicity, and their results are compared to that in standard MI and other simple approaches to missing data. The aims of these two case studies are as follows.

1. To assess the plausibility of the MAR assumption for ethnicity data in UK primary care databases, and;
2. To examine the association between ethnicity and the prevalence of type 2 diabetes diagnoses in UK primary care databases.

Chapter 7 concludes this thesis by discussing the methodological development of calibrated-δ adjustment MI and weighted MI, highlighting the implications of these methods, and identifying potential areas for future work.

2

Missing data and multiple imputation

2.1 INTRODUCTION

This chapter presents an overview of the issues raised by missing data in medical research and the available methods for handling missing data. Section 2.2 outlines the different missingness patterns and missingness mechanisms which are the key concepts underpinning the analysis of incomplete data. Section 2.3 reviews simple methods for handling missing data, including complete record analysis and single value imputation which serves as a platform for the ideas of multiple imputation. Section 2.4 gives an introduction to multiple imputation and various aspects relevant to multiple imputation analysis, including Rubin's rules [20, 21] for multiple imputation inference, methods for performing univariate and multivariate multiple imputation, key considerations in specifying the imputation model, and the use of multiple imputation under different assumptions about the underlying missingness mechanism.

Missing data refer to values which were intended to be recorded in a study, but for some reason were not [25].

Missing data are commonly seen in medical research, where data are often missing due to non-response. Carpenter and Plewis [26] categorised non-response into four different types which are presented below, with relevant examples in the context of primary healthcare.

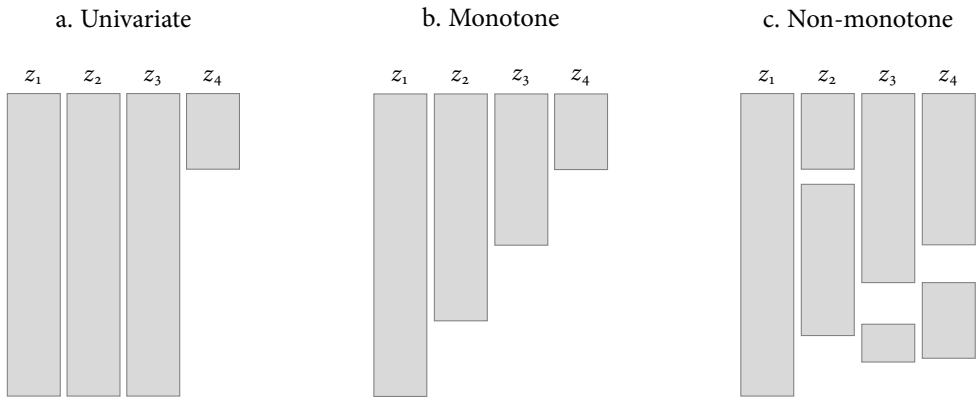
1. Unit non-response. Individuals fail to provide enough information for the response to be deemed usable. In primary care, unit non-response is present when individuals register with their general practices but do not consult, and therefore very limited information about them is observed;
2. Item non-response. Data from individuals are partially observed, i.e. they have at least some observed data. In primary care, some individuals might, for example, have their blood pressure measured at regular intervals, but not their cholesterol level. Recording of health indicators in primary care may be influenced by pay-for-performance initiatives, see section 6.2.3 for more details;
3. Wave non-response. In longitudinal studies, some sample members fail to participate in a particular wave of the study. In primary care, data are not collected at fixed time points and as a result, individuals often have gaps between consultations during which no information is recorded;
4. Attrition. In longitudinal studies, some initially cooperative sample members drop out of the study before the study ends. In primary care, individuals either die or leave their general practices, from which point no further information is recorded.

Note that when data are referred to as 'missing', their values are not recorded but are assumed to exist. For example, a living person would have blood pressure irrespective of whether the blood pressure measurements are recorded. On the other hand, missingness due to death is a fundamentally different concept, in that the missing values cannot usually be said to exist. This type of missingness is not the focus of this thesis.

With rectangular datasets, data are arranged in such a way that the rows correspond to individuals and the columns correspond to variables, and there are three main classes of overall missingness patterns [17] (figure 2.1). Let $z = (z_{ij})$ denote the $n \times p$ matrix containing data on p variables for n individuals in the dataset, where z_{ij} is the value of variable z_j for the ith individual. With missing values, define an $n \times p$ matrix of the response indicator $r = (r_{ij})$, such that $r_{ij} = 1$ if z_{ij} is observed, and $r_{ij} = 0$ otherwise. The matrix r stores the locations of missing observations in z, and therefore defines the pattern of missing data. The three main classes of missingness patterns are as follows.

1. Univariate missingness pattern. Missing values occur in a single variable (figure 2.1a);
2. Monotone missingness pattern. Missing values occur in more than one variable, and the variables can be ordered such that for individual i, $z_{j+1}, \ldots, z_p$ are missing whenever z_j is missing (figure 2.1b). Attrition is an example of the monotone missingness pattern. Here, repeated measurements will be missing for an individual in all subsequent waves of a study, following the individual's drop-out from the study;
3. Non-monotone missingness pattern. Missing values occur in more than one variable and

Figure 2.1. Schematic representation of different missingness patterns with four variables.



* Note: the unshaded areas represent missing values.

there is a random scatter of missing values across the variables (figure 2.1c), such as in the case of item or wave non-response.

While missingness patterns indicate which values are missing and observed in a given dataset, missingness mechanisms describe the relation between the probability of the data being missing and the values of the data, both observed and missing. This relation can be characterised by the missing data model $p(r \mid z, \alpha)$ of the response indicator r given the data z, where α is the unknown parameters [27]. Let z^{obs} and z^{mis} collectively denote the observed and missing values of z. Taken together, $z = (z^{\text{obs}}, z^{\text{mis}})$ contain data for the complete dataset. Rubin [28] defined three broad classes of missingness mechanisms, each with distinct implications for the analysis of partially observed data:

1. Missing completely at random (MCAR). The probability of data being missing does not depend on any observed or unobserved information, $p(r \mid z^{\text{obs}}, z^{\text{mis}}, \alpha) = p(r \mid \alpha)$. For example, a laboratory blood sample is MCAR if it is accidentally dropped because the chance of this random event occurring is the same for all individuals, regardless of their lifestyle factors or health outcomes. Under MCAR, the complete records (i.e. individuals with complete data) are a random sample of the set of originally identified individuals. Therefore, it can be assumed that missing data are similar to the observed data;
2. Missing at random (MAR). The probability of data being missing does not depend on unobserved information, conditional on observed information, $p(r \mid z^{\text{obs}}, z^{\text{mis}}, \alpha) = p(r \mid z^{\text{obs}}, \alpha)$. Data can be considered MAR given the observed groups, where observations within the same group have the same probability of being missing. For example, blood pressure measurements are MAR conditional on sex, if women are more likely than men to have their blood pressure measured and sex is fully observed. In other words, within the same sex, the distribution of blood pressure values is the same, whether or not blood pressure is measured;
3. Missing not at random (MNAR). The probability of data being missing depends on unobserved information, such as the missing values or an unmeasured variable, $p(r \mid z^{\text{obs}}, z^{\text{mis}}, \alpha)$ does not simplify. For example, blood pressure measurements are MNAR if individuals with high blood pressure are more likely to have their blood pressure measured, even after controlling for other fully observed factors such as age and sex.

It might be possible to identify whether missing data are MCAR or not in incomplete datasets, for example, by using χ^2 tests or logistic regression models for the binary response indicator of the incomplete variable, conditional on other fully observed variables [29]. Conversely, the observed data alone are not sufficient to distinguish between MAR and MNAR mechanisms. Further external information or untestable assumptions are required to describe the relationship between missingness and the unseen values which are not available in the observed data [17]. For inference under MNAR, a joint model for z_{mis} and r is required. However, there will typically be a wide range of models for r that match the data observed, and these models may result in very different inferences. Therefore, assessing the robustness of results to potential departure from the posited assumption about the missingness mechanism by conducting supplemental sensitivity analyses under alternative plausible assumptions should play a central role [30, 31].

The following sections describe available methods for handling missing data and their validity under different assumptions about the underlying missingness mechanism.

2.3 SIMPLE METHODS FOR HANDLING MISSING DATA

There are many approaches for dealing with missing data, the choice of which one to use depends on both the missingness pattern and the assumption about the missingness mechanism. The goal of any procedure for dealing with missing data is to retain the characteristics of the data and the association between variables, in order to obtain valid and efficient inferences [17]. Methods for handling missing data should generally be evaluated based on the following criteria [17, 32]. First, the method should yield unbiased parameter estimates over a wide range of parameters. Second, the estimated standard errors should be close to the true standard deviations of the parameter estimates, i.e. confidence intervals should cover the true values of the parameters with probability close to the nominal level, implying an accurate probability of Type I error. Third, once bias and standard errors have been addressed, the method should yield precise estimates with small standard errors and narrow confidence intervals, which lead to a lower Type II error and increased power.

Several simple, or 'ad-hoc', approaches were proposed to overcome the issue of missing data; they are often resorted to in practice, mainly for computational convenience rather than for their validity. Despite their ease of implementation, these approaches generally require more restrictive assumptions about missing data that rarely hold in practice, and their shortcomings have been discussed extensively in the literature [16, 33–37]. Section 2.3.1 discusses complete record analysis, one of the most frequently adopted approaches for handling missing data which is usually performed before other more sophisticated methods are considered. As demonstrated in this section, complete record analysis can provide valid inferences in certain settings where the missingness mechanism assumption required for the validity of other model-based methods is violated. Section 2.3.2 then presents two single imputation methods which serve as a platform for the ideas of multiple imputation (section 2.4).

2.3.1 *Complete record analysis*

A complete record analysis (CRA) is an analysis restricted to the complete records, where individuals with missing values in one or more variables are excluded from the analysis. It is the default method for accommodating missing data in most statistical packages.

Under the MCAR mechanism, the subset of complete records represents a random, though smaller than originally intended, sample of individuals in the study. Therefore, results from a CRA will be unbiased. The validity of CRA is also extended to certain settings beyond MCAR; one of which is where the analysis consists of fitting a regression model for some outcome variable on one or several covariates. Results from CRA are valid if the probability of being a complete record is independent of the outcome when conditioned on the covariates, regardless of whether missingness occurs in the outcome or the covariates [27, 38–40]. As an example, in standard cohort studies where individuals are followed-up over time from study entry, it might be reasonable to assume that missingness in the covariates measured at baseline is not caused by the outcome, which is measured later after a period of follow-up [40]. This is because the future values of the outcome are yet to be determined at the time data for baseline covariates are collected. Under this assumption, CRA can provide unbiased estimates of association.

However, even when this assumption holds and CRA is unbiased, the method is not optimally efficient since it involves discarding information from individuals with partially observed data. In multivariable analyses, a relatively small number of missing values in each variable may cause a large proportion of the study sample to be excluded, which can lead to reduced efficiency and lower power.

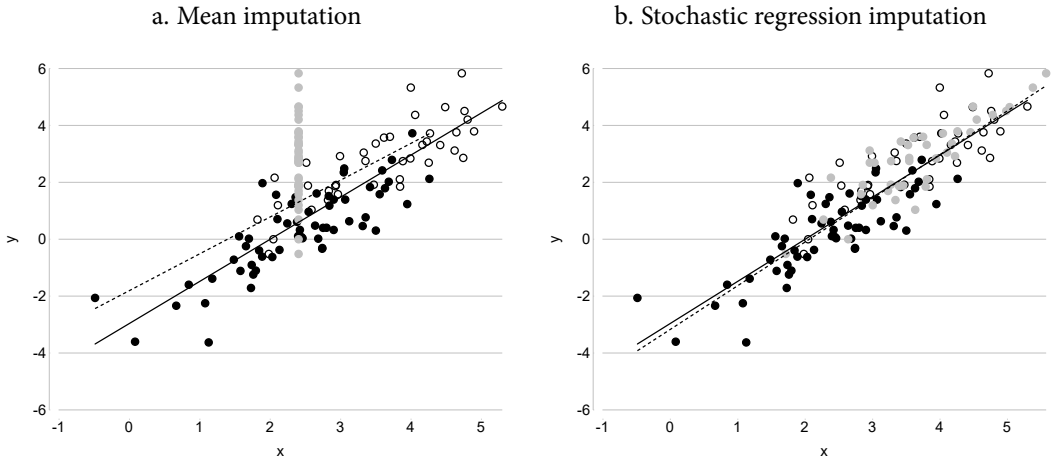
To avoid the potential loss of information associated with CRA, it is sensible to seek ways of drawing on information available in the incomplete records. For settings where CRA provides valid inferences such as described above, Bartlett et al. [40] proposed an augmented complete record approach. This approach improves efficiency through the specification of an additional model for the probability of missingness conditional on fully observed variables. Alternatively, when some individuals have partially observed information, rather than excluding these individuals entirely, it is tempting to 'fill in' their missing values with plausible values and proceed with the analysis as planned. There are several single-value imputation techniques which impute the missing values in order to yield a single completed dataset for analysis. The following sections discuss two single imputation approaches which motivate the ideas of multiple imputation.

2.3.2 *Single-value imputation*

One of the most common single imputation approaches is mean imputation, where missing values in a variable are replaced by the marginal mean of the observed values of that variable. This method has many disadvantages. First, it ignores the relationship between the incomplete variable and other variables in the analysis, which can lead to biased estimates of association. Second, although this method maintains the original sample size and is easy to implement, the variability in the data is reduced since all missing values are replaced with the same value. Consequently, variance tends to be underestimated.

In a simple example, let y and x denote two continuous variables whose joint distribution is bivariate normal and the full-data analysis is a linear regression of y on x, with some values of

Figure 2.2. Single-value imputation: linear associations between the outcome variable y (complete) and the covariate x (MAR conditional on y), when x is imputed using mean imputation and stochastic regression imputation.



* Note: black lines: regression lines in the full data; dotted lines: regression lines in completed data after single imputation of x; black circles: observed values; hollow circles: missing values; grey circles: imputed values.

x missing. Figure 2.2a shows the result of a simple simulation with 100 observations of y and x, where 40% of the values in x are assumed to be MAR conditional on y. Missing values in x are imputed using mean imputation, which biases the slope of the regression line toward 0. The variability in the observed data is also not reflected in the imputed values in mean imputation.

Stochastic regression imputation was proposed to preserve the association between variables in the observed data and correct bias created by the reduced variability in mean imputation [37]. In this method, a regression *imputation model* is fitted to the complete records, where the dependent variable is the incomplete variable and the independent variables are other fully observed variables which are predictive of the incomplete variable. Predicted values are obtained from the imputation model and are augmented with a residual term to replace missing values. The residual term is normally distributed with mean 0 and variance equal to the mean square error of the imputation model. This incorporates the variability in the observed data into the imputed values, resulting in more plausible standard errors compared to mean imputation. In the above example, each missing value in x is replaced with $\dot{x}_i = \hat{\theta}_0 + \hat{\theta}_y y_i + \dot{\varepsilon}_i$, where $\dot{\varepsilon}_i$ is a random draw from the normal distribution $\mathrm{N}(0, \hat{\sigma}^2)$ and $\hat{\sigma}^2$ is the mean square error of the imputation model $x_i = \theta_0 + \theta_y y_i + \varepsilon_i; \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ (figure 2.2b). By comparing figures 2.2a and 2.2b, it is clear that stochastic regression imputation produces more plausible imputed values compared to mean imputation.

Although stochastic regression imputation is clearly preferred to mean imputation, the analysis model generates standard errors of parameter estimates which are still generally too small. This is because the uncertainty in estimating the θ parameters and σ of the imputation model is not acknowledged. Indeed, the method can be modified to be fully stochastic, with an extra step that draws the values of the imputation model's parameters from their distributions; in fact this is a key step in multiple imputation which is presented in section 2.4. However, this process still does not account for the fact that missing values are replaced with reasonable guesses, and then data are analysed as if there are no guesses. In other words, the originally

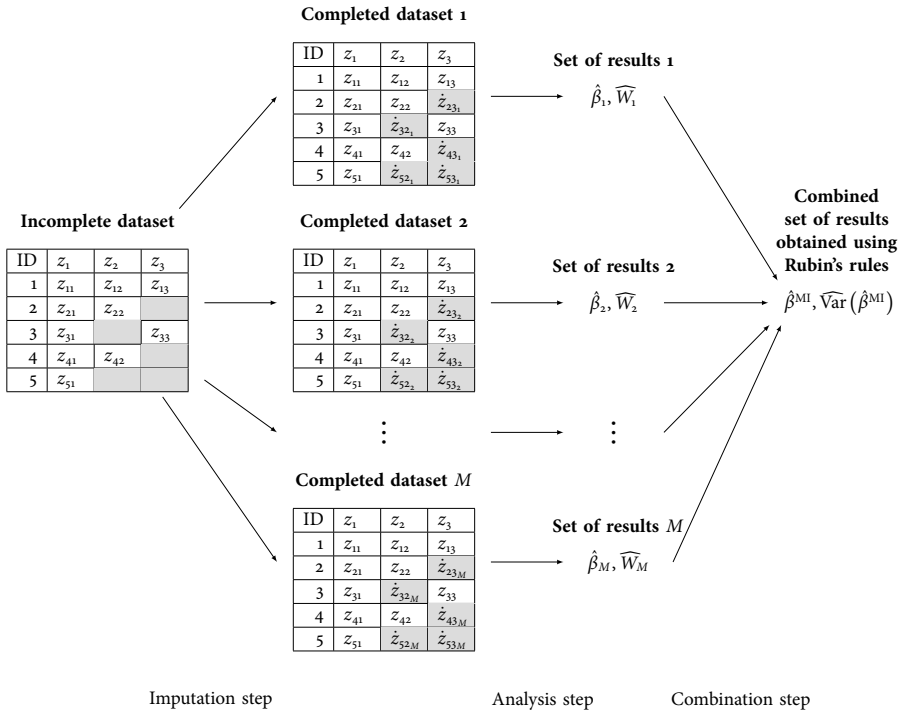
observed data are not distinguished from the imputed data. This implies that standard errors of the analysis model's parameters are generally underestimated, resulting in confidence intervals that are too narrow [17]. The next section introduces multiple imputation which builds on the single imputation methods presented above.

2.4 MULTIPLE IMPUTATION

Multiple imputation (MI) [20] is a popular approach for the analysis of partially observed data, which was originally developed as a technique for handling survey non-response. Since its inception in the 1970s, MI has increasingly been regarded as the standard method for handling missing data in medical research [18]. The number of publications which applied MI to impute missing data in the main analysis or explored new methodological extensions and adaptations of the method has grown exponentially [37, 41]. MI has been incorporated in standard statistical software packages [42–44], making it more accessible to medical researchers and enabling a wider application of the method in recent years. MI is a suitable tool for addressing item missingness which is commonly seen in medical research [25], where each individual has at least some observed data.

The aim of MI is to obtain valid inferences in the presence of missing data [45]. Similar to single imputation, MI retains all individuals in the analysis, but missing data are imputed in such a way that fully accounts for the uncertainty about them. Analysis with missing data using MI generally proceeds in three steps, as illustrated in figure 2.3.

Figure 2.3. Schematic representation of multiple imputation analysis.



* Note: $\hat{\beta}_m$ and $\widehat{W}_m$: estimate of parameter β and its variance obtained from the mth completed dataset, respectively, $m = 1, \ldots, M$; $\hat{\beta}^{\mathrm{MI}}$ and $\widehat{V}\left(\hat{\beta}^{\mathrm{MI}}\right)$: combined parameter estimate and associated variance for inference.

1. Imputation step. Fill in missing values $M > 1$ times with plausible values generated from an imputation model, which is the Bayesian posterior predictive distribution of missing data conditional on the observed data, to create M completed datasets;
2. Analysis step. Perform the analysis that would have been used in the absence of missing data identically in each completed dataset, storing parameter estimates of interest and associated standard errors;
3. Combination step. Use Rubin's rules [20, 21] (section 2.4.1) to combine results from the M completed datasets into a single set of parameter estimates and standard errors, accounting for the variability in results across the completed datasets and reflecting the uncertainty about the missing values.

The next section discusses Rubin's rules [20, 21] which are used in the last step of MI analysis to obtain the combined set of results for inference.

2.4.1 *Rubin's rules for multiple imputation inference*

Let $\hat{\beta}_m$ denote the estimate of parameter β obtained from perfoming the main analysis in the mth completed dataset, and $\widehat{W}_m$ denote its estimated variance; $m = 1, \ldots, M$. The MI estimator of β is given by

$$\hat{\beta}^{\mathrm{MI}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m.$$

The associated total variance estimator is expressed as

$$\widehat{\mathrm{Var}}\left(\hat{\beta}^{\mathrm{MI}}\right) = \left(1 + \frac{1}{M}\right) \widehat{B} + \widehat{W},$$

where

$$\widehat{B} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta}_m - \hat{\beta}^{\mathrm{MI}}\right)^2 \text{ and } \widehat{W} = \frac{1}{M} \sum_{m=1}^{M} \widehat{W}_m$$

denote the between-imputation and within-imputation variances, respectively, and $\left(1 + \frac{1}{M}\right)$ is an adjustment factor which reflects the extra variability as a consequence of using a finite number of imputations instead of an infinite number of imputations.

Inference for $\hat{\beta}_{\mathrm{MI}}$, including hypothesis tests and confidence intervals, is based on the t-distribution with ν degrees of freedom, where

$$\frac{\hat{\beta}^{\mathrm{MI}} - \beta_{H_0}}{\sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}^{\mathrm{MI}}\right)}} \sim t_\nu, \quad \text{and } \nu = (M-1)\left[1 + \frac{\widehat{W}}{\left(1 + \frac{1}{M}\right)\widehat{B}}\right]^2.$$

For a finite number of imputations, the fraction of missing information (FMI) for β which is a measure of the loss of precision due to missing data, is estimated by

$$\mathrm{FMI} = \frac{\left(1 + \frac{1}{M}\right)\frac{\widehat{B}}{\widehat{W}} + 2/(\nu+3)}{\left(1 + \frac{1}{M}\right)\frac{\widehat{B}}{\widehat{W}} + 1} \approx \frac{\widehat{B}}{\widehat{W} + \widehat{B}},$$

where $\left(1 + \frac{1}{M}\right)\frac{\widehat{B}}{\widehat{W}}$ is the relative increase in variance (RVI) due to missing data.

The relative efficiency (RE) [20] of using M imputations versus an infinite number of imputations is given by

$$\mathrm{RE} = \left(1 + \frac{\mathrm{FMI}}{M}\right)^{-1}. \tag{2.1}$$

Rubin's rules [20, 21] are generic and can be applied to estimators that are normally distributed [46]. Some statistics such as odds ratios or hazard ratios require sensible transformation before combination, while statistics that are not estimators such as p-values or likelihood ratio test statistics cannot be combined using these rules [47]. Despite missing values being imputed from a Bayesian model, MI using Rubin's rules [20, 21] can provide valid frequentist inferences with asymptotically unbiased point estimates and variance estimation such that confidence intervals achieve their nominal coverage [20, 27].

As explained in the next section, the standard implementation of MI provides valid inferences based on the ignorability assumption with an underlying MAR mechanism. It is also possible, although more complex, to perform MI under the MNAR assumption for missing data (see section 2.5).

2.4.2 *The Bayesian justification of multiple imputation*

Rubin [20] recommends that imputations are drawn from a Bayesian posterior predictive distribution of missing data. Thus, MI provides an approximation to a fully Bayesian procedure. Schafer [48] coined the term 'Bayesianly proper' to define an imputation procedure in which missing values are imputed by independent draws from the posterior predictive distribution of a Bayesian model, under a parametric model for the complete data (and, if necessary, a model for the missingness mechanism) and a prior distribution of the unknown model parameters. This process reflects uncertainty in the imputation, including errors in the predicted values and estimation errors in the parameters of the imputation model [48]. This section provides a broad overview of the Bayesian justification of MI, including how the posterior predictive distribution of missing data is derived and how imputed values are simulated from this distribution.

The joint distribution of the full data can be written as $p\left(z^{\text{obs}}, z^{\text{mis}}, r \mid \beta, \alpha\right)$, which depends on parameters β for the data z that are of interest, and parameters α for the response indicator r that are seldom of interest. When z contains missing values, this joint model cannot be evaluated in the normal way. However, the distribution of the observed data can be obtained by integrating out the missing data as follows

$$p\left(z^{\text{obs}}, r \mid \beta, \alpha\right) = \int p\left(r \mid z^{\text{obs}}, z^{\text{mis}}, \beta, \alpha\right) p\left(z^{\text{obs}}, z^{\text{mis}} \mid \beta, \alpha\right) dz^{\text{mis}}. \tag{2.2}$$

Bayesian inference about β and α is based on the observed-data posterior distribution of β and α, which combines the observed-data likelihood with a prior distribution

$$\begin{aligned} p\left(\beta, \alpha \mid z^{\text{obs}}, r\right) &\propto p\left(\beta, \alpha\right) L\left(\beta, \alpha \mid z^{\text{obs}}, r\right), \\ \text{where} \quad L\left(\beta, \alpha \mid z^{\text{obs}}, r\right) &\propto p\left(z^{\text{obs}}, r \mid \beta, \alpha\right). \end{aligned} \tag{2.3}$$

Given that (i) β and α are *a priori* independent, i.e. $p\left(\beta, \alpha\right) = p\left(\beta\right) p\left(\alpha\right)$, and (ii) z_{mis} is MAR conditional on z_{obs}, (2.2) can be written as

$$p(z^{\text{obs}}, r \mid \beta, \alpha) = p(r \mid z^{\text{obs}}, \alpha) p(z^{\text{obs}} \mid \beta).$$

It follows that β and α are *a posteriori* independent [27], i.e. (2.3) becomes

$$\begin{aligned} p\left(\beta, \alpha \mid z^{\text{obs}}, r\right) &\propto p\left(\beta\right) L\left(\beta \mid z^{\text{obs}}\right) p\left(\alpha\right) L\left(\alpha \mid z^{\text{obs}}, r\right) \\ &\propto p\left(\beta \mid z^{\text{obs}}\right) p\left(\alpha \mid z^{\text{obs}}, r\right). \end{aligned}$$

Bayesian inference about β can therefore be based on the observed-data posterior distribution $p\left(\beta \mid z^{\text{obs}}\right)$, ignoring the missing data mechanism. Furthermore, the missing data problem can be separated from the main analysis, such that

$$p\left(\beta \mid z^{\text{obs}}\right) = \int p\left(\beta \mid z^{\text{mis}}, z^{\text{obs}}\right) p\left(z^{\text{mis}} \mid z^{\text{obs}}\right) dz^{\text{mis}}, \tag{2.4}$$

i.e. the first term in the integral of (2.4) represents the posterior distribution of β given the complete data, which is the Bayesian version of the main analysis [49].

In MI analysis, (2.4) is approximated in two steps, first by imputing z^{mis} M times from the posterior predictive distribution $p\left(z^{\text{mis}} \mid z^{\text{obs}}\right)$, i.e. the *imputation* step, followed by evaluating the complete-data posterior distribution at each of the imputed values z_m^{mis}, i.e. the *complete-data analysis* step. Rubin's rules [20, 21] approximate the integral by summarising over the M draws of z^{mis},

$$p\left(\beta \mid z^{\text{obs}}\right) \approx \frac{1}{M} \sum_{m=1}^{M} p\left(\beta \mid z^{\text{obs}}, z_m^{\text{mis}}\right).$$

Imputing missing data from the posterior predictive distribution $p\left(z^{\text{mis}} \mid z^{\text{obs}}\right)$ requires the specification of a parametric imputation model $p\left(z^{\text{obs}}, z^{\text{mis}} \mid \theta\right)$ with a prior distribution of θ and the computation of conditional distributions for drawing z^{mis} from this model, since

$$p\left(z^{\text{mis}} \mid z^{\text{obs}}\right) = \int p\left(z^{\text{mis}} \mid z^{\text{obs}}, \theta\right) p\left(\theta \mid z^{\text{obs}}\right) d\theta.$$

An imputation of z^{mis} can therefore be created by simulating a random draw of the unknown parameters from their observed-data posterior, $\dot{\theta} \sim p\left(\theta \mid z^{\text{obs}}\right)$, followed by a random draw of the missing values from the posterior predictive distribution, $\dot{z} \sim p\left(z^{\text{mis}} \mid z^{\text{obs}}, \dot{\theta}\right)$ [50].

While drawing from $p\left(z^{\text{mis}} \mid z^{\text{obs}}, \theta\right)$ is generally straightforward, sampling from the distribution $p\left(\theta \mid z^{\text{obs}}\right)$ is often not simple. The next section discusses methods for performing sampling from the posterior predictive distribution.

2.4.3 *Univariate and multivariate multiple imputation*

In univariate MI of a single incomplete variable, a regression imputation model for the incomplete variable conditional on other fully observed variables is fitted to the complete records. The imputation model can be tailored to the variable being imputed; for example, a linear regression model for a continuous variable or a logistic regression model for a binary variable. Let $\hat{\theta}$ and $\widehat{\text{Var}}\left(\hat{\theta}\right)$ denote the estimated parameters of the imputation model and their covariance matrix. Imputations are simulated from the posterior predictive distribution of the incomplete variable using $\dot{\theta}$ and the appropriate probability distribution, where $\dot{\theta}$ is randomly drawn from the posterior distribution commonly approximated by the multivariate normal distribution, $\dot{\theta} \sim \text{MVN}\left(\hat{\theta}, \widehat{\text{Var}}\left(\hat{\theta}\right)\right)$ [20, 47].

The example in section 2.3.2, where (x, y) follow a bivariate normal distribution and x has some missing values, is now used to demonstrate how to obtain proper imputation in a simple setting. Suppose x is MCAR or MAR conditional on y, let $y = (1, y)'$ and n^{obs} denote the number of subjects with observed x. To obtain proper imputation for x [47], the linear regression model $x \mid y \sim \text{N}\left(\theta_0 + \theta_y y, \sigma^2\right)$ is fitted to n^{obs} subjects whose x values are observed. Let $\hat{\theta}$ denote the vector of estimated parameters with covariance matrix $\widehat{\text{Var}}\left(\hat{\theta}\right)$ and root mean square error $\hat{\sigma}$. A draw of $\dot{\theta}$ and $\dot{\sigma}$ is obtained from their joint posterior distribution assuming a non-informative

prior [20]. First, $\dot{\sigma}$ is drawn as

$$\dot{\sigma} = \hat{\sigma}\sqrt{\frac{n^{\text{obs}} - 2}{e}},$$

where 2 is the number of parameters to be estimated in the imputation model for x and e represents a random draw from the χ^2 distribution with $n^{\text{obs}} - 2$ degrees of freedom. This is followed by a draw of $\dot{\theta}$ as

$$\dot{\theta} = \hat{\theta} + \frac{\dot{\sigma}}{\hat{\sigma}} u_1 \widehat{\text{Var}}\left(\hat{\theta}\right)^{\frac{1}{2}},$$

where u_1 is a row vector of 2 independent random draws from the standard normal distribution and $\widehat{\text{Var}}\left(\hat{\theta}\right)^{\frac{1}{2}}$ is the Cholesky decomposition of $\widehat{\text{Var}}\left(\hat{\theta}\right)$. For individual i with missing x_i, each missing value is imputed with

$$\dot{x}_i = \dot{\theta} y_i + u_{2i}\dot{\sigma},$$

where u_{2i} is a random draw from the standard normal distribution. This procedure is repeated M times, creating M completed datasets. Similarly, univariate MI procedures for binary, unordered, and ordered categorical variables are described in White et al. [47], section 2.

For multivariate missing data, when the missingness pattern is monotone distinct [51] (parameters of the univariate conditional models have independent priors), incomplete variables can be arranged in a monotone pattern with increasing amounts of missing values. They are then imputed without iteration by sequential specification of univariate models conditional on the complete and previously imputed variables [51].

When the missingness pattern is non-monotone, factorisation of the joint likelihood function into independent likelihood functions for the incomplete variables might become impossible [48]. As a consequence, imputation cannot be performed from independent univariate models as in the case of the monotone missingness pattern. One approach for dealing with the non-monotone missingness pattern is to assume a joint parametric model for the data and draw imputed values from the resulting distribution of the missing data given the observed data. A popular choice for the parametric joint model is the multivariate normal model, although as noted in section 2.4.2, direct simulation from the corresponding predictive distribution $p\left(z^{\text{mis}} \mid z^{\text{obs}}\right)$ of missing data given the observed data is not simple. This is due to an intrinsic dependence of the unknown parameter θ on the posterior distribution $p\left(\theta \mid z^{\text{obs}}\right)$. Assuming an underlying multivariate normal model, the data augmentation method which is an iterative Bayesian Markov chain Monte Carlo (MCMC) procedure can be used to approximate the distribution of missing data conditional on the observed data [48, 52]. In data augmentation, the observed data z_{obs} are augmented with unobserved data z^{mis}, such that the full-data posterior $p\left(\theta \mid z^{\text{obs}}, z^{\text{mis}}\right)$ is easier to draw from. At the tth iteration, a draw of $z^{\text{mis}(t)}$ is generated from the distribution $p\left(z^{\text{mis}} \mid z^{\text{obs}}, \theta^{(t-1)}\right)$ conditional on the observed data and the previous draw of θ, i.e. the I(mputation)-step. This is followed by a new draw of $\theta^{(t)}$ from its posterior distribution $p\left(\theta \mid z^{\text{obs}}, z^{\text{mis}(t)}\right)$ given the augmented data, i.e. the P(osterior)-step. Iterating between these two steps sets up a Markov chain that converges to a stationary distribution, which is the joint distribution $p\left(z^{\text{mis}}, \theta \mid z^{\text{obs}}\right)$ of missing data and parameters given the observed data. Proper imputations can be obtained by running this process for a large number of iterations and storing the results of several I-steps with enough iterations in between to ensure independence. The draws of $\theta^{(t)}$ approximate the distribution $p\left(\theta \mid z^{\text{obs}}\right)$;

likewise, the draws of $z^{\text{mis}(t)}$ correspond to the distribution $p\left(z^{\text{mis}} \mid z^{\text{obs}}\right)$.

In practice, however, defining a multivariate joint model for the data can often be a challenging task. For example, difficulty arises when the incomplete variables are of different types, e.g. continuous, binary, unordered and ordered categorical. This makes the specification of conventional models, such as the multivariate normal model, theoretically inappropriate. Similarly, the relationship between variables can be complex, e.g. non-linear [53]. A more practical approach to joint modelling for multivariate missing data was introduced by van Buuren et al. [23]. It is commonly referred to as fully conditional specification (FCS) [53] or multivariate imputation by chained equations (MICE) [43]. MICE involves specifying a series of univariate models for the conditional distribution of each partially observed variable, given other variables. For q incomplete variables, instead of defining a q-variate joint distribution, MICE proceeds in an iterative fashion as follows.

1. Fill in missing values in each of the incomplete variables with randomly chosen observed values of that variable;
2. For each incomplete variable, discard its first filled-in values, define a univariate regression imputation model for that variable conditional on all other variables, and replace the missing values with random draws from this conditional model;
3. Repeat step 2 for each of the q incomplete variables in turn, completing one iteration;
4. Repeat step 3 to create a few iterations until convergence is attained.

MICE provides a flexible alternative to joint modelling imputation since the imputation task is split into specifying relatively simple univariate conditional models that are more conventional, e.g. linear regression for continuous variables, logistic regression for binary variables, and so on. However, a known theoretical downside of MICE is that the assumed underlying joint distribution of the conditional models may not always exist; in other words, the conditional models may be incompatible [47, 51, 53]. Two conditional models are said to be incompatible if there exists no joint model for which the conditionals for the corresponding variables equal these conditional models [40]. One consequence of incompatibility of the conditional models is that the distribution of the imputed values, and hence the results of the analysis, may differ depending on the order in which the variables are updated in the chain equations sampler. This is also known as the 'order effects' [54]. However, it was shown that under a set of linear regression conditional models with all other variables as covariates and no interactions, MICE is equivalent to a Gibbs sampler, drawing from a multivariate normal distribution [51]. For three incomplete binary variables, MICE under a set of logistic regression conditional models with all other variables included as main effects only is equivalent to joint modelling imputation under a log-linear model with the three-way factor term set to 0 [51]. Under sufficient conditions (including compatibility of the conditional models) given by Liu et al. [55], the stationary distribution of the Markov chain generated in MICE (assuming that this stationary distribution exists and the chain converges to it) converges to the posterior predictive distribution of missing data. This corresponds to a joint Bayesian model as the sample size tends to infinity. Hughes et al. [54] and Liu et al. [55] independently provided an additional non-informative margins condition, according to which the imputed values yielded from MICE and joint modelling correspond to the same predictive distribution. This condition requires that, together with compatibility of the conditional models

and assuming the Markov chain generated in MICE converges to a stationary distribution, the joint prior distribution factorises into independent priors. In a simulation study, Hughes et al. [54] examined the consequences for MICE when compatibility of the conditional models holds but the non-informative margins condition is not satisfied. The order effects were found to to be present, but their average magnitude was small and did not cause bias [54]. This supports previous findings from van Buuren et al. [53] and van Buuren [51] that MICE appears robust to incompatibility of the conditional models.

2.4.4 *Specifying the imputation model*

In MI analysis, the missing data problem (the imputation step) is separated from the complete-data analysis (the analysis step). This separation can be advantageous but can also lead to difficulties in specifying the imputation model. In particular, for MI of incomplete covariates, the imputation model might be incompatible with the analysis model in the sense described in section 2.4.3, and this can lead to biased parameter estimates in the main analysis [40]. Liu et al. [55] distinguished between the two departures from compatibility, as described below.

1. Semi-compatibility. The imputation model can be made compatible with the analysis model by restricting some parameters in the imputation model to 0;
2. Incompatibility. The imputation model cannot be made compatible with the analysis model by restricting some parameters in the imputation model to 0.

Semi-compatibility implies that the analysis model is a restricted version of the imputation model and all features in the analysis model are preserved in the imputation model. In practice, this implies that there are some considerations for choosing the imputation models to ensure (semi-)compatibility. Most importantly, all variables in the analysis model, including the outcome variable, must be present in the imputation model [47, 56]. For example, if the outcome is not appropriately included in the imputation model for the incomplete covariate, its association with the covariate will be biased towards 0 in the main analysis. This is because imputations are created assuming the incomplete covariate is independent of the outcome [57]. This issue is illustrated in the analysis of the QRISK tool for cardiovascular risk prediction using primary care data [14]. After MI had been used for handling missing values in the analysis, cardiovascular risk was found to be surprisingly unrelated to cholesterol (coded as the ratio of the total to high density lipoprotein cholesterol). Updated results were later obtained following a revision of the imputation procedure which showed a clear association between cholesterol ratio and cardiovascular risk. A possible explanation for this change in results is that one component of the survival outcome, which comprises the time-to-event and the event indicator, was omitted in the original imputation model. In survival analysis where the outcome is assumed to follow the Cox proportional hazards model, White and Royston [58] investigated different functions of time-to-event. It was showed that by including both the Nelson–Aalen estimator of the cumulative hazard function and the event indicator in the imputation model for imputing missing covariate values, the results obtained were less biased compared to using the log survival time [58].

MI allows for the inclusion of auxiliary variables in the imputation model. These variables are not in the analysis model but provide information about the missing values and/or the missingness mechanism [46, 47, 59]. Good candidates for auxiliary variables are variables which

are predictive of both the missing values and the probability of data being missing. Including these auxiliary variables in the imputation model will improve the plausibility of the MAR assumption and reduce bias. Auxiliary variables that are only predictive of the missing values can help to reduce the standard errors of estimates in the analysis model. Conversely, variables that are only predictive of the probability of data being missing will not add information and should not be included in the imputation model [46].

When imputed values are generated from an imputation model which is more restricted than the analysis model, incompatibility arises as a result of the imputation model containing more assumptions than the analysis model. The implication of this incompatibility on inference depends on the plausibility of the extra assumptions made by the imputation model. If these extra assumptions are plausible, such imputation is unbiased and 'super-efficient', i.e. Rubin's variance estimator is positively biased and confidence intervals have coverage that is greater than the nominal level [45, 60, 61]. If, however, these extra assumptions do not hold, MI may lead to biased estimates.

The imputation model should also reflect any structure in the analysis model, such as interactions and non-linearity. If there is an interaction between one incomplete covariate to be imputed and another covariate which is fully observed and discrete, the dataset can be stratified according to the values of the discrete covariate. Imputation can then be performed separately in each stratum in the usual way, without having to explicitly incorporate the interaction into the imputation model [56]. Alternatively, passive imputation and 'just another variable' (JAV) imputation [56] were proposed for handling interactions and multiplicative terms. However, Seaman et al. [62] reported that while JAV gives consistent estimation for linear regression with a quadratic or interaction term under MCAR, the method may be biased when data are MAR. JAV can also lead to severe bias when used for the logistic regression [62].

Recently, Bartlett et al. [40] proposed the substantive model compatible fully conditional specification (SMC-FCS) method. SMC-FCS provides modification to the normal FCS procedure to ensure that incomplete covariates are imputed from models which are compatible with the analysis model. Under the MAR assumption, the method was shown to give consistent estimates for a range of analysis models, including those with non-linear covariate effects or interactions, provided that the imputation models are compatible and correctly specified.

When an incomplete covariate in the analysis model is a ratio, e.g. the Body Mass Index or cholesterol ratio, imputing the numerator and denominator of the ratio separately can yield implausible values of the ratio. Returning to the QRISK analysis example [14], another possible explanation for the original results showing no association between cholesterol ratio and cardiovascular risk is that total and high density lipoprotein cholesterol values were imputed separately rather than as a ratio. Specifically, missing data in high density lipoprotein cholesterol were imputed with values close to 0, resulting in very large values of the ratio. For such situations, Morris et al. [63] recommended imputing the ratio either directly, or passively by imputing the log-transformed numerator and denominator and then deriving the ratio.

To account for potential departure from the MAR assumption, the next section describes methods for analysing incomplete data under the MNAR assumption and discusses how such methods can be used for performing sensitivity analyses in the presence of missing data.

Strategies for handling missing data under the MNAR assumption generally involves defining a model for the missingness mechanism, which describes how missingness depends on both observed and unobserved information. This implies that in practice, it is necessary to posit a model for either the association between the probability of observing a variable and its unseen values; or the different joint distributions of observed and missing data across the patterns of missing observations. Due to the potential complexity of modelling the MNAR mechanism, analyses assuming data are MNAR are relatively infrequently performed and reported in the applied literature [41].

There are two main approaches for specifically addressing missing data under the MNAR mechanism; the pattern-mixture model [64, 65] and the selection model [24, 27, 66]. These approaches are based on two different factorisations of the joint distribution of the data which comprise the variables and the response indicators. A detailed discussion on pattern-mixture and selection modelling approaches is provided in Carpenter and Kenward [46], chapter 10.

2.5.1 *Pattern-mixture models*

The pattern-mixture factorisation of the joint distribution of the data allows for direct modification of the behaviour of missing data. MI is therefore well-suited for this purpose, such that the imputation process can be directly intervened to reflect potential departure from the MAR assumption. Below is an illustrative example of a general approach to pattern-mixture modelling using MI, in which missing values are imputed from a missing data distribution which differs for each missingness pattern. Consider a linear regression model for an outcome y conditional on covariates $x = \left(x_1, \ldots, x_p\right)'$, where y is incomplete and x is fully observed. For each individual i, $i = 1, \ldots, n$, let r_i denote the response indicator taking values 1 if y_i is observed, and 0 otherwise. Under the MAR assumption, the distribution of y conditional on x is assumed to be the same, whether or not y is observed

$$p(y_i \mid x_i, r_i = 0) = p(y_i \mid x_i, r_i = 1).$$

Therefore, in the standard implementation of MI assuming data are MAR, the distribution $p(y \mid x)$ is estimated in the complete records and then used to impute missing values in the incomplete records. When the missingness mechanism is MNAR, it follows that

$$p(y_i \mid x_i, r_i = 0) \neq p(y_i \mid x_i, r_i = 1).$$

Thus, MI under the MNAR assumption can proceed by first creating several imputed values assuming data are MAR, followed by changing the imputations by some chosen fixed quantity δ. This δ is also known as the 'sensitivity parameter' and represents key differences in the conditional distribution between the observed and missing data ($\delta = 0$ under MAR). Having imputed the missing values, the analysis model is fitted to each completed dataset and the results are combined using Rubin's rules [20, 21] in the usual way. For example, suppose that in an analysis, the variable blood pressure contains some missing values. The MAR assumption means that conditional on other fully observed variables in the analysis, the distributions of observed and missing blood pressure values are the same. Suppose that individuals with missing blood pressure values have,

on average, measurements that are 10 mmHg below the values predicted assuming data are MAR. Under this additional assumption about departure from MAR, missing blood pressure measurements can first be imputed using standard MI assuming data are MAR, followed by subtracting $\delta = 10$ mmHg from the imputed values. Rubin's rules [20, 21] can then be used for inference and the results can be compared to that under no deviation from the MAR assumption.

This is a simple example of the pattern-mixture model, such that the completed dataset represents a mixture of potentially different imputations for different missingness patterns. Valid inference is obtained if δ is chosen correctly. However, the suitability of the chosen value of δ cannot be validated in the observed data. As a result, a sensitivity analysis is often performed in which the analysis is repeated for different choices of δ. This approach is relatively straightforward to implement and communicate, and expert opinions about the plausible values for the sensitivity parameter (if available) can be directly incorporated in the imputation.

2.5.2 *Selection models*

Following the selection factorisation of the joint distribution of the data, the selection modelling approach involves explicitly defining a missingness model for the probability of data being missing, which is estimated jointly with the analysis model. Fully Bayesian analyses are particularly convenient for this purpose.

Heckman [24] proposed a sample selection model, also known as the Heckman model, which was successfully applied to account for data being MNAR in incomplete outcome variables. This method deals with selected samples by defining two linear regression models for the outcome variable and the response indicator, which are joint by their error terms. Continuing with the above example of a linear regression model and assuming that data are only missing in y, the analysis model is given by

$$\begin{aligned} y_i &= x_i\beta + \varepsilon_{1i}; \\ \varepsilon_{1i} &\overset{\text{iid}}{\sim} N(0, \sigma^2_{\varepsilon_1}). \end{aligned} \tag{2.5}$$

Let r denote the response indicator of y, such that $r_i = 1$ if y_i is observed, and 0 otherwise. A selection model representing the non-random sampling of the missingness process is defined as

$$p\left(r_i = 1 \mid x_i^h\right) = \Phi\left(x_i^h\alpha\right), \tag{2.6}$$

where α is a vector of the unknown parameters, x^h denotes variables which are thought to be predictive of missingness in y (x^h may partly or fully contain variables in x), and Φ denotes the cumulative distribution function of the standard normal distribution. The selection equation is defined through a latent normally distributed variable z, such that

$$\begin{aligned} z_i &= x_i^h\alpha + \varepsilon_{2i}; \\ \varepsilon_{2i} &\overset{\text{iid}}{\sim} N(0, \sigma^2_{\varepsilon_2}). \end{aligned} \tag{2.7}$$

This model is related to the analysis model through the error term ε_{2i}, with $r_i = 1$ if $z_i >= 0$ and $r_i = 0$ if $z_i < 0$. Hence, the Heckman model is defined jointly by (2.5) and (2.7). The joint

distribution of the error term is a bivariate normal distribution

$$\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix} \quad , \quad \begin{pmatrix}\sigma^2_{\varepsilon_1} & \rho\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}\\ \rho\sigma_{\varepsilon_2}\sigma_{\varepsilon_1} & \sigma^2_{\varepsilon_2}\end{pmatrix}\right],$$

where ρ is the correlation coefficient between ε_1 and ε_2, and $\sigma_{\varepsilon_2} = 1$ under the probit link in (2.6).

It also follows that

$$E\left(z \mid y\right) = z\alpha + \rho\frac{\varepsilon_2}{\varepsilon_1}\left(y - x\beta\right).$$

Under the MAR assumption, $\rho = 0$; thus, the selection and analysis equations can be estimated separately. When data are MNAR, $\rho \neq 0$; these equations must therefore be estimated jointly. The strength of the MNAR mechanism increases with increasing values of ρ.

To obtain unbiased estimates of the β parameters in the analysis model, Heckman [24] proposed the following two-step procedure.

1. Estimate parameters α in (2.6) by maximum likelihood. These estimates are used to compute estimates of the inverse Mills ratio (IMR) as

$$\widehat{\mathrm{IMR}}_i = \frac{\phi\left(x_i\beta\right)}{\Phi\left(x_i^h\alpha\right)},$$

   where ϕ and Φ are the probability density and cumulative distribution functions of the standard normal distribution, respectively.
2. Include the estimated Mill ratios as an additional covariate in the analysis model and obtain estimates of the analysis model's parameters as follows

$$y_i = x_i\beta + \widehat{\mathrm{IMR}}_i\beta_{\mathrm{IMR}} + \varepsilon_{3i},$$
$$\varepsilon_{3i} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\varepsilon_3}).$$

In general, this approach can be inconsistent in small samples. Particularly, if there is a significant overlap between x and x^h, or if they are identical, issues with identification can arise [67]. An example based on the set-up of the Heckman model is presented in section 5.3.

As an alternative to the probit link which is used to define the selection process in the Heckman model, the logit link can also be used to define the selection model. Using the logit link, the selection model in the above example becomes

$$\mathrm{logit}\left[p\left(r_i = 1 \mid x_i, y_i\right)\right] = \alpha_0 + \alpha_x x_i + \delta y_i,$$

where δ represents the (fixed) selection sensitivity parameter, which is the adjusted log odds ratio relating the chance of observing the outcome to its underlying unseen values. A value of $\delta = 0$ implies a MAR mechanism for the missing data. As a result, the analysis and selection models may be fitted separately. A value of $\delta \neq 0$ represents departure towards the MNAR mechanism. Therefore, the analysis and selection models must be fitted simultaneously in a Bayesian framework, using software like WinBUGS [68].

It is often hard to choose values of δ; plausible values of δ can be elicited by drawing on expert insights [69]. A sensitivity analysis exploring a range of different values for δ can be performed to assess the sensitivity of results to potential departure towards MNAR assumptions [70].

Carpenter et al. [66] proposed a re-weighting method which uses importance weighting to approximate a selection model. In this method, MI is used to impute missing data under the MAR assumption, and the analysis model is fitted to each completed dataset. In the final step of

MI, instead of averaging the results for the imputation estimates using Rubin's rules [20, 21], a weighted average is performed, up-weighting the imputation estimates that are more plausible under the assumed MNAR mechanism. This re-weighting method was recently critiqued by Hayati Rezvan et al. [71], who reported that while this method outperformed standard MI under the MNAR mechanisms considered, it still provided biased parameter estimates even when a large number of imputations was used.

2.6 SUMMARY

This chapter presents an overview of the issues raised by missing data and the various methods proposed for the analysis of incomplete data. The concepts covered in this chapter are therefore essential for understanding the theory and results presented in subsequent chapters of this thesis. In particular, this chapter gives a detailed description of MI of missing data, including the MI procedure, the standard implementation of MI under the MAR assumption, and key considerations when specifying the imputation model. The last part of this chapter discusses available methods for handling missing data under the assumption of data being MNAR.

The next chapter introduces and evaluates weighted multiple imputation, one of the two population-calibrated multiple imputation approaches proposed in this thesis.

3

Weighted multiple imputation of a binary covariate when the outcome variable is binary

Multiple imputation (MI), as introduced in the last chapter, is a model-based alternative to simple methods for handling missing data. MI is increasingly regarded as the standard procedure for the analysis of partially observed data in medical research [18]. In practice, MI is commonly implemented under the missing at random (MAR) assumption; that is, the probability of data being missing does not depend on unobserved information, conditional on observed information. However, for missing data generated under missing not at random (MNAR) mechanisms, the standard implementation of MI assuming data are MAR might not be satisfactory.

For an incomplete variable in a given dataset, its population-level marginal distribution might be available in an external data source. Assume that the study sample is expected to be representative of the population in terms of the incomplete variable. If standard MI under the MAR assumption yields a post-imputation marginal distribution of the incomplete variable that does not agree with its known population distribution, then standard MI might not be the appropriate approach for handling missing data, probably due to a potential MNAR mechanism. Therefore, standard MI can potentially be improved by matching the incomplete variable's post-imputation distribution to the population level. As highlighted in chapter 1, the broad aim of this thesis is to explore the use of the incomplete variable's population distribution in the imputation process in order to improve standard MI under general MNAR mechanisms.

This chapter proposes and evaluates the *weighted multiple imputation* method for utilising external information containing the incomplete variable's population distribution in MI to calibrate inference to the population. Throughout this thesis, the focus is on incomplete binary/categorical variables that are included as covariates in the analysis model of interest. In particular, this chapter explores the setting of missing data in an incomplete binary covariate of an analysis model, when the outcome variable is also binary.

Section 3.2 describes the steps in weighted MI and explains how standard MI can be augmented with appropriately calculated probability weights. These weights are derived using the incomplete variable's population distribution to closely match the post-imputation distribution to the population level.

Weighted MI is then evaluated in a univariate missing data setting where missingness occurs in a single covariate, under missingness mechanisms of increasing complexity. In particular, section 3.3 presents an analytic study of a 2×2 contingency table with a fully observed outcome variable and a single partially observed covariate. Bias in the analysis model's parameter estimates are derived analytically and compared between standard MI and weighted MI under different missingness mechanisms for the covariate. Sections 3.4 and 3.5 feature univariate simulation studies of the same setting to investigate other finite-sample properties of weighted MI.

Section 3.6 investigates a multivariate missing data setting where missingness occurs in more than one covariate. Weighted MI is evaluated and compared to standard MI and complete record analysis in multivariate simulation studies of a three-way contingency table. These studies feature a fully observed outcome variable and two partially observed covariates, and different missingness mechanisms are used for generating missing values in the covariates.

In weighted MI of an incomplete binary/categorical variable, the complete records are assigned weights which are calculated based on the incomplete variable's population marginal distribution taken from an external dataset. The derivation of the marginal and conditional weights is introduced in sections 3.2.1 and 3.2.2.

The procedure of the proposed weighted MI method for an incomplete binary/categorical variable is as follows. In the imputation step, weights derived from the population marginal distribution of the incomplete variable are attached to the complete records, and a weighted (multinomial) logistic regression model is fitted to the complete records to obtain the maximum likelihood estimates of the imputation model's parameters, $\hat{\theta}$, and their asymptotic sampling variance, $\widehat{\mathrm{Var}}\left(\hat{\theta}\right)$. New parameters are then drawn from the large-sample normal approximation of their posterior distribution, $\mathrm{N}\left(\hat{\theta}, \widehat{\mathrm{Var}}\left(\hat{\theta}\right)\right)$, assuming non-informative priors. Finally, imputed values are drawn from the (multinomial) logistic regression model using these newly drawn parameters. Note that no weights are used when fitting the analysis model to the completed data, since the 'fixing' has been done in the imputed data. This is consistent with Rubin's MI philosophy in which the imputation step is separated from the analysis step [20].

The rationale for the derivation of the marginal and conditional weights used in weighted MI of an incomplete binary/categorical variable is outlined next.

3.2.1 *Derivation of the marginal weights*

The idea of using weights in MI is related to the technique of post-stratification weights, which is commonly used to deal with survey non-response when the population distribution/totals of some of the variables are known [72]. To post-stratify the sample, weights are calculated to bring the sample distribution in line with the population. Suppose that in a survey, one of the variables measured is ethnicity, which is categorised into four groups as White, Black, Asian, and Other. If the population distribution of ethnicity is available, the distribution of ethnicity among survey respondents can be compared with the population distribution. Suppose that a proportion $p^{\mathrm{obs}} = 0.8$ of the survey respondents give their ethnicity as White, where as the population has $p^{\mathrm{pop}} = 0.6$ in this category. The White category is over-represented in the survey respondents, but can be made representative of the population by assigning to the White respondents a post-stratification weight $w^{\mathrm{ps}} < 1$ such that

$$w^{\mathrm{ps}} = 1/\left(p^{\mathrm{obs}}/p^{\mathrm{pop}}\right) = 1/\left(0.8/0.6\right) = 0.75.$$

A discussion on post-stratification weighting can be found in Raghunathan [72], chapter 2. In adapting this idea to MI, it is necessary to address the complication arising because the completed data obtained after MI consist of both observed and imputed (missing) data.

Naive use of post-stratification weights in MI will recover the correct population distribution in the imputed data, but not when combined with the observed data. Since the observed data remain the same, the distribution in the completed (observed and imputed) data will not be matched to that in the population. Therefore, some compensation for the lack of representation in the observed data is needed in the imputed data, so that the correct population distribution can be recovered after imputation. Continuing with the survey example, suppose that 200 individuals

are surveyed, 100 of whom respond with information about their ethnicity. A proportion $p^{\text{obs}} = 0.8$ of these 100 respondents is in the White group. If the population proportion of this group is $p^{\text{pop}} = 0.6$, the survey sample is expected to contain 120 White individuals. This implies that among the 100 individuals with incomplete ethnicity, missing values in 40 individuals need to be imputed as White, i.e. the proportion of the White ethnic group required in the missing data, p^{req}, is equal to 0.4. To make the completed data in this category representative of the population, respondents of this category need to be weighted in the imputation model by

$$1/\left(p^{\text{obs}}/p^{\text{req}}\right) = 1/(0.8/0.4) = 0.5.$$

This weight is smaller than the corresponding naive post-stratification weight, since it compensates for the over-representation among the respondents of White ethnicity.

More generally, suppose that data of an L-level variable x are collected for a sample of size n, resulting in x being observed for n^{obs} subjects and missing for n^{mis} subjects, $n^{\text{obs}} + n^{\text{mis}} = n$. Let p_l^{obs} and p_l^{req} denote the level-l proportion of x in the observed and imputed data, respectively, such that $p_l^{\text{obs}} n^{\text{obs}} = n_l^{\text{obs}}$, and $p_l^{\text{req}} n^{\text{mis}} = n_l^{\text{req}}$, $l = 1, \ldots, L$. Let p_l^{pop} denote the level-l proportion of x in the population, which is assumed to be known from an external dataset. The aim here is to find p_l^{req} for each level of x such that the number of subjects in the completed data after imputation is equal to the expected number implied by the corresponding population proportion, $p_l^{\text{pop}} n = n_l^{\text{obs}} + n_l^{\text{req}}$. The level-$l$ proportion of x required in the imputed data, p_l^{req}, is estimated from the following

$$\begin{aligned} & p_l^{\text{pop}} n = p_l^{\text{obs}} n^{\text{obs}} + p_l^{\text{req}} n^{\text{mis}}; \\ \rightarrow\ & p_l^{\text{req}} = \frac{p_l^{\text{pop}} n - p_l^{\text{obs}} n^{\text{obs}}}{n^{\text{mis}}}. \end{aligned} \tag{3.1}$$

The weight for group l, which is referred to as the *marginal* weight and denoted by w_l^{m}, is

$$w_l^{\text{m}} = 1/(p_l^{\text{obs}}/p_l^{\text{req}}).$$

3.2.2 *Derivation of the conditional weights*

The marginal weights introduced above only depend on the population distribution of the incomplete variable. However, if there are (fully observed) covariates in the imputation model, the associations of these variables with the incomplete variable's distribution are not reflected in such weights. This is demonstrated in an example of a 2×2 contingency table in section 3.3, where marginal weighted MI does not recover the correct distribution of the incomplete covariate x when missingness in x depends on the outcome variable y.

To address this, the marginal weights are adjusted to obtain another set of weights, termed the *conditional* weights, which account for covariates in the imputation model. These weights are derived using the marginal distribution of the incomplete variable obtained after having estimated the parameters of an imputation model (under the MAR assumption) in the complete records. In other words, instead of deriving the weights using the incomplete variable's observed distribution, its predicted distribution in completed data yielded by standard MI is weighted against the population distribution.

Suppose that an imputation model is fitted to the complete records, and the corresponding predicted probabilities of the incomplete variable (averaged over the covariates) are obtained

and applied to the missing data. Let p_l^{pred} denote the resulting predicted level-l proportion of x in the completed data, then the level-l proportion required in the imputed data is given by

$$p_l^{\text{req}} = \frac{p_l^{\text{pop}} n - p_l^{\text{pred}} n^{\text{obs}}}{n^{\text{mis}}}, \tag{3.2}$$

and the conditional weight for group l, denoted by w_l^{c} , is

$$w_l^{\text{c}} = 1/(p_l^{\text{pred}}/p_l^{\text{req}}).$$

In this approach, the effects of covariates in the imputation model are reflected in the predicted probabilities p_l^{pred}, which are then used to derive the conditional weights for weighted MI. Note that p^{pred} is obtained once in the observed data, and used across the imputations.

Weights can become non-positive when the numerators of (3.1) and (3.2) are non-positive. Non-positive weights can be attributed to the following reasons.

1. Sampling variation can result in the number of subjects in group l observed in the sample to be slightly higher than what is implied by the reference proportion;
2. Some subjects who in truth do not belong to group l are misclassified into this group in the sample, which can happen due to errors in data recording or when the incomplete variable contains many similar groups;
3. The MNAR mechanism is such that data are only ever missing in one category, which is also accompanied by sampling variation. For example, suppose the level-l proportion of x in the population is 0.3, which implies that in a sample of 1007 subjects, 302 of them are expected to be in this group. However, due to random sampling, there are in fact 304 subjects sampled to be in group l, among whom x is only missing for one subject, i.e. $n_l^{\text{obs}} > p_l^{\text{pop}} N$. As a result, the weight for this category will be negative, but very close to 0;
4. The sample is not representative of the reference population in terms of x.

Within the context of this thesis, only positive probability weights are used in weighted MI, as a negative weight implies that some subjects should be 'removed' from the observed data. Therefore, negative weights can be set to a very small positive value close to 0, and the corresponding groups get imputed very infrequently. However, when a category is markedly over-represented in the observed data, this means that it may remain over-represented even if no missing values are imputed to this category. In such situations, the proposed weighted MI procedure cannot produce completed datasets whose incomplete variable distribution exactly matches the population level.

To implement marginal and conditional weighted MI of an incomplete binary/categorical variable, I have written and released two Stata commands, `mi impute wlogit/wmlogit` [73] (available on SSC). These are based on the Stata community-contributed command `uvis` [74] which performs univariate MI of an incomplete variable.

The next section describes an analytic study which aims to explore the settings in which marginal and conditional weighted MI can successfully remove bias in a 2×2 contingency table.

3.3 Analytic study – bias calculation in a 2×2 contingency table

This analytic study is conducted to analytically explore bias in weighted MI in a univariate missing data setting, where missingness occurs in a single covariate in the analysis model. A

working example of a 2×2 contingency table with a fully observed binary outcome variable y and a partially observed binary covariate x is chosen. Bias in the analysis model's parameter estimates is derived analytically for standard MI (assuming data are MAR) and marginal and conditional weighted MI under several missingness mechanisms for x.

This exercise allows for a comparison between complete record analysis, standard MI, and marginal and conditional weighted MI in terms of bias; and for identifying potential scenarios where using weighted MI to accommodate missing values in the covariate is preferred to the standard MI approach. The results here apply directly to higher dimensional contingency tables with one partially observed variable, but the algebra is considerably more complex.

3.3.1 *Method*

Suppose it is of interest to study the association between a binary covariate x whose levels are indexed by l and a binary outcome y whose levels are indexed by k; l and k take values 0 or 1. The full-data distribution of x and y (table 3.1) is assumed to be identical to the population distribution, such that the population marginal distribution of x is given by $p_l^{\text{pop}} = \frac{n_{l+}}{n_{++}}$.

The analysis model is

$$\text{logit}\left[p\left(y = 1 \mid x\right)\right] = \beta_0 + \beta_x x, \tag{3.3}$$

whose parameters can be written in terms of cell counts as

$$\beta_0 = \ln\left(\frac{n_{01}}{n_{00}}\right), \beta_x = \ln\left(\frac{n_{11}n_{00}}{n_{01}n_{10}}\right).$$

In addition, suppose that some values of x are set to missing. Let r be the response indicator taking values 1 if x is observed, and 0 if x is missing. Four missingness mechanisms are considered for x in this study; these are all possible missingness mechanisms for missing values in x in this setting (excluding an interaction between x and y in the selection model for x). Missingness mechanism assumptions for x as well as the corresponding selection models and cell-wise probabilities of observing x are presented in table 3.2. Observed cell count n_{lk}^{obs} can be written as a product of the full-data cell count n_{lk} and the cell-wise probability $p\left(r_{lk} = 1\right)$ of observing x, such that $n_{lk}^{\text{obs}} = n_{lk}p\left(r_{lk} = 1\right)$.

Missing values in x are imputed using standard MI (under the MAR assumption) and marginal and conditional weighted MI, after which the β parameters in the analysis model (3.3) are estimated, and bias defined as $\hat{\beta} - \beta$ is calculated.

In standard MI, the imputation model

$$\text{logit}\left[p\left(x = 1 \mid y\right)\right] = \theta_0 + \theta_y y$$

Table 3.1. Analytic study: distribution of x and y in the full data.

	$y = 0$	$y = 1$	$\sum_{y=0}^{1}$
$x = 0$	n_{00}	n_{01}	n_{0+}
$x = 1$	n_{10}	n_{11}	n_{1+}
$\sum_{x=0}^{1}$	n_{+0}	n_{+1}	n_{++}

Table 3.2. Analytic study: models for missingness in x.

Linear predictor of selection model $\text{logit}\left[p\left[(r=1 \mid x, y)\right]\right.$	Selection probability $p\left(r_{lk}=1\right)$	Label
α_0	p_r	M1
$\alpha_0 + \alpha_y y$	p_{r_k}	M2
$\alpha_0 + \alpha_x x$	p_{r_l}	M3
$\alpha_0 + \alpha_x x + \alpha_y y$	$p_{r_{lk}}$	M4

* Note: r: response indicator of x; l and k: index categories of x and y, respectively; l, k take values 0 or 1.

is fitted to the n_{++}^{obs} complete records to obtain the θ estimates, such that

$$\hat{\theta}_0^{\text{s}} = \ln\left(\frac{n_{10}^{\text{obs}}}{n_{00}^{\text{obs}}}\right), \hat{\theta}_y^{\text{s}} = \ln\left(\frac{n_{11}^{\text{obs}} n_{00}^{\text{obs}}}{n_{01}^{\text{obs}} n_{10}^{\text{obs}}}\right), \hat{p}_{lk}^{\text{s}} = \frac{n_{lk}^{\text{obs}}}{n_{+k}^{\text{obs}}}, \tag{3.4}$$

where $\hat{p}_{lk}^{\text{s}}$ denotes the predicted probability of $x = l$, given $y = k$ in the complete records; l and k take values 0 or 1. In weighted MI, the same imputation model is fitted to the *weighted* complete records, $n_{lk}^{\text{obs}} w_l^{\text{m/c}}$, where a marginal/conditional weight $w_l^{\text{m/c}}$ is assigned to subjects with observed $x = l$. Parameter estimates and predicted probabilities of the weighted imputation model are

$$\hat{\theta}_0^{\text{m/c}} = \ln\left(\frac{n_{10}^{\text{obs}} w_1^{\text{m/c}}}{n_{00}^{\text{obs}} w_0^{\text{m/c}}}\right); \hat{\theta}_y^{\text{m/c}} = \ln\left(\frac{n_{11}^{\text{obs}} n_{00}^{\text{obs}}}{n_{01}^{\text{obs}} n_{10}^{\text{obs}}}\right); \hat{p}_{lk}^{\text{m/c}} = \frac{n_{lk}^{\text{obs}} w_l^{\text{m/c}}}{\sum_{l=0}^{1} n_{lk}^{\text{obs}} w_l^{\text{m/c}}}. \tag{3.5}$$

In marginal weighted MI, the level-l proportion of x required among the imputed values, p_l^{req}, and the weight, w_l^{m}, are

$$p_l^{\text{req}} = \frac{n_{l+} - n_{l+}^{\text{obs}}}{n_{++}^{\text{mis}}};$$
$$w_l^{\text{m}} = \frac{n_{l+} - n_{l+}^{\text{obs}}}{n_{++}^{\text{mis}}} \cdot \frac{n_{++}^{\text{obs}}}{n_{l+}^{\text{obs}}}.$$

In conditional weighted MI, these become

$$p_l^{\text{req}} = \frac{n_{l+} - p_l^{\text{pred}} n_{++}^{\text{obs}}}{n_{++}^{\text{mis}}};$$
$$w_l^{\text{c}} = \frac{n_{l+} - p_l^{\text{pred}} n_{++}^{\text{obs}}}{n_{++}^{\text{mis}}} \cdot \frac{1}{p_l^{\text{pred}}},$$

where

$$p_l^{\text{pred}} = \frac{n_{l+}^{\text{obs}} + \sum_{k=0}^{1} \hat{p}_{lk}^{\text{s}} n_{+k}^{\text{mis}}}{n_{++}}.$$

3.3.2 *Analytic calculations*

Full analytic calculations of bias in standard MI and marginal and conditional weighted MI under missingness models M1–M4 are presented below.

Let $\hat{n}_{lk}$ denote the estimated count for cell (l, k), such that $\hat{n}_{lk} = n_{lk}\hat{b}_{lk}$, where $\hat{b}_{lk}$ denotes the estimated cell-wise bias. When $\hat{b}_{lk} \neq 1$, bias in the estimates of model (3.3)'s parameters can

generally be written as

$$\text{Bias}(\hat{\beta}_0) = \ln\left(\frac{n_{01}\hat{b}_{01}}{n_{00}\hat{b}_{00}}\right) - \ln\left(\frac{n_{01}}{n_{00}}\right) = \ln\left(\frac{\hat{b}_{01}}{\hat{b}_{00}}\right);$$
$$\text{Bias}(\hat{\beta}_x) = \ln\left(\frac{n_{00}\hat{b}_{00}n_{11}\hat{b}_{11}}{n_{10}\hat{b}_{10}n_{01}\hat{b}_{01}}\right) - \ln\left(\frac{n_{00}n_{11}}{n_{10}n_{01}}\right) = \ln\left(\frac{\hat{b}_{00}\hat{b}_{11}}{\hat{b}_{10}\hat{b}_{01}}\right).$$

1. M1 – bias when x is MCAR

   The cell-wise probability of observing X is the same for all cells, $p\left(r_{lk}=1\right)=p_r$. Standard MI leads to unbiased estimates of cell counts as shown below

$$\begin{aligned}\hat{n}^{\text{s}}_{lk} &= n^{\text{obs}}_{lk} + n^{\text{req}}_{lk}\\ &= n^{\text{obs}}_{lk} + n^{\text{mis}}_{+k}\frac{n^{\text{obs}}_{lk}}{n^{\text{obs}}_{+k}}\\ &= n^{\text{obs}}_{lk}\frac{n_{+k}}{n^{\text{obs}}_{+k}}\\ &= n_{lk}p_r\frac{1}{p_r}\\ &= n_{lk},\end{aligned}$$

   which implies that the estimates of model (3.3)'s parameters are unbiased, as expected.

   The marginal weight, w^{m}_l, in weighted MI is given by

$$w^{\text{m}}_l = \frac{n_{l+}(1-p_r)}{n_{++}(1-p_r)}\cdot\frac{n_{++}p_r}{n_{j+}p_r} = 1.$$

   Thus, marginal weighted MI is equivalent to standard MI and also provides unbiased parameter estimates of model (3.3).

   Since estimated cell counts are unbiased in standard MI, the level-l predicted probability of x in the completed data, p^{pred}_l, is also unbiased, $p^{\text{pred}}_l = \frac{n_{l+}}{n_{++}}$. The conditional weight, w^{c}_l, is

$$w^{\text{c}}_l = \frac{n_{l+} - \frac{n_{l+}}{n_{++}}n_{++}p_r}{n_{++}(1-p_r)}\cdot\frac{n_{++}}{n_{l+}} = 1,$$

   which implies that conditional weighted MI also provides unbiased estimates of the analysis model's parameters.

2. M2 – bias when x is MAR conditional on y

   The cell-wise probability of observing x is dependent on y, $p\left(r_{lk}=1\right)=p_{r_k}$. Since standard MI gives unbiased estimates of cell counts as shown below, parameter estimates of model (3.3) are also unbiased in standard MI,

$$\begin{aligned}\hat{n}^{\text{s}}_{lk} &= n^{\text{obs}}_{lk} + n^{\text{req}}_{lk}\\ &= n^{\text{obs}}_{lk} + n^{\text{mis}}_{+k}\frac{n^{\text{obs}}_{lk}}{n^{\text{obs}}_{+k}}\\ &= n^{\text{obs}}_{lk}\frac{n_{+k}}{n^{\text{obs}}_{+k}}\\ &= n_{lk}p_{r_k}\frac{1}{p_{r_k}}\\ &= n_{lk}.\end{aligned}$$

Estimated count for cell (l, k) in marginal weighted MI, $\hat{n}^{\mathrm{m}}_{lk}$, is biased as follows

$$\begin{aligned}
\hat{n}^{\mathrm{m}}_{lk} &= n^{\mathrm{obs}}_{lk} + n^{\mathrm{req}}_{lk} \\
&= n_{lk} p_{r_k} + n^{\mathrm{mis}}_{+k} \frac{n_{lk} w^{\mathrm{m}}_{l}}{\sum_{l=0}^{1} n_{lk} w^{\mathrm{m}}_{l}} \\
&= n_{lk} \left(p_{r_k} + \frac{n^{\mathrm{mis}}_{+k} w^{\mathrm{m}}_{l}}{\sum_{l=0}^{1} n_{lk} w^{\mathrm{m}}_{l}} \right) \\
&= n_{lk} \hat{b}^{\mathrm{m}}_{lk},
\end{aligned}$$

which implies that marginal weighted MI produces biased parameter estimates of model (3.3). Since $p^{\mathrm{pred}}_{l} = \frac{n_{l+}}{n_{++}}$ in standard MI, conditional weights in weighted MI can be shown to be equal to 1, as

$$\begin{aligned}
w^{\mathrm{c}}_{l} &= \frac{n_{l+} - p^{\mathrm{pred}}_{l} n^{\mathrm{obs}}_{++}}{n^{\mathrm{mis}}_{++}} \cdot \frac{n_{++}}{n_{l+}} \\
&= \frac{n_{l+} - \frac{n_{l+}}{n_{++}} n^{\mathrm{obs}}_{++}}{n^{\mathrm{mis}}_{++}} \cdot \frac{n_{++}}{n_{l+}} \\
&= \frac{n_{l+} \frac{n^{\mathrm{mis}}_{++}}{n_{++}}}{n^{\mathrm{mis}}_{++}} \cdot \frac{n_{++}}{n_{l+}} \\
&= 1.
\end{aligned}$$

Hence, conditional weighted MI is equivalent to standard MI, and is unbiased for model (3.3)'s parameter estimates.

3. M3 – bias when x is MNAR dependent on x

The cell-wise probability of observing x is dependent on x, $p\left(r_{lk} = 1\right) = p_{r_l}$. Standard MI produces biased estimates of cell counts as shown below

$$\begin{aligned}
\hat{n}^{\mathrm{s}}_{lk} &= n^{\mathrm{obs}}_{lk} + n^{\mathrm{req}}_{lk} \\
&= n^{\mathrm{obs}}_{lk} + n^{\mathrm{mis}}_{+k} \frac{n^{\mathrm{obs}}_{lk}}{n^{\mathrm{obs}}_{+k}} \\
&= n^{\mathrm{obs}}_{lk} \frac{n_{+k}}{n^{\mathrm{obs}}_{+k}} \\
&= n_{lk} \left(\frac{p_{r_l} \sum_{l=0}^{1} n_{lk}}{\sum_{l=0}^{1} n_{lk} p_{r_l}} \right) \\
&= n_{lk} \hat{b}^{\mathrm{s}}_{lk},
\end{aligned}$$

which leads to a biased estimate of β_0 but an unbiased estimate of β_x, since

$$\begin{aligned}
\ln\left(\frac{\hat{b}^{s}_{00} \hat{b}^{s}_{11}}{\hat{b}^{s}_{10} \hat{b}^{s}_{01}} \right) &= \ln\left(\frac{\frac{p_{r_0} \sum_{l=0}^{1} n_{l0}}{\sum_{l=0}^{1} n_{l0} p_{r_l}} \cdot \frac{p_{r_1} \sum_{l=0}^{1} n_{l1}}{\sum_{l=0}^{1} n_{l1} p_{r_l}}}{\frac{p_{r_1} \sum_{l=0}^{1} n_{l0}}{\sum_{l=0}^{1} n_{l0} p_{r_l}} \cdot \frac{p_{r_0} \sum_{l=0}^{1} n_{l1}}{\sum_{l=0}^{1} n_{l1} p_{r_l}}} \right) \\
&= \ln\left(1\right) \\
&= 0.
\end{aligned}$$

In marginal weighted MI, the marginal weights are given by

$$w_{l} = \frac{n^{\mathrm{mis}}_{l+}}{n^{\mathrm{mis}}_{++}} \cdot \frac{n^{\mathrm{obs}}_{++}}{n^{\mathrm{obs}}_{l+}}$$

$$= \frac{n_{l+}\left(1-p_{r_l}\right)}{\sum_{l=0}^{1} n_{l+}\left(1-p_{r_l}\right)} \cdot \frac{\sum_{l=0}^{1} n_{l+}p_{r_l}}{n_{l+}p_{r_l}},$$

from which the ratio of the two marginal weights for $x = 1$ and $x = 0$ can be written as

$$\frac{w_1^{\mathrm{m}}}{w_0^{\mathrm{m}}} = \frac{p_{r_0}\left(1-p_{r_1}\right)}{p_{r_1}\left(1-p_{r_0}\right)}.$$

For $l = 1$, estimated count for cell $(1, k)$ is unbiased under marginal weighted MI as follows

$$\begin{aligned}
\hat{n}_{1k}^{\mathrm{m}} &= n_{1k}^{\mathrm{obs}} + n_{1k}^{\mathrm{req}} \\
&= n_{1k}p_{r_1} + n_{+k}^{\mathrm{mis}} \frac{n_{1k}p_{r_1}w_1^{\mathrm{m}}}{n_{0k}p_{r_0}w_0^{\mathrm{m}} + n_{1k}p_{r_1}w_1^{\mathrm{m}}} \\
&= n_{1k}\left(p_{r_1} + \frac{n_{+k}^{\mathrm{mis}}p_{r_1}w_1^{\mathrm{m}}}{n_{0k}p_{r_0}w_0^{\mathrm{m}} + n_{1k}p_{r_1}w_1^{\mathrm{m}}}\right) \\
&= n_{1k}\left(p_{r_1} + \frac{n_{+k}^{\mathrm{mis}}p_{r_1}\frac{w_1^{\mathrm{m}}}{w_0^{\mathrm{m}}}}{n_{0k}p_{r_0} + n_{1k}p_{r_1}\frac{w_1^{\mathrm{m}}}{w_0^{\mathrm{m}}}}\right) \\
&= n_{1k}\left\{p_{r_1} + \frac{\left[n_{0k}\left(1-p_{r_0}\right) + n_{1k}\left(1-p_{r_1}\right)\right]\frac{p_{r_0}\left(1-p_{r_1}\right)}{\left(1-p_{r_0}\right)}}{n_{0k}p_{r_0} + n_{1k}\frac{p_{r_0}\left(1-p_{r_1}\right)}{\left(1-p_{r_0}\right)}}\right\} \\
&= n_{1k}\left\{p_{r_1} + \frac{\left[n_{0k}\left(1-p_{r_0}\right) + n_{1k}\left(1-p_{r_1}\right)\right]p_{r_0}\left(1-p_{r_1}\right)}{\left[n_{0k}\left(1-p_{r_0}\right) + n_{1k}\left(1-p_{r_1}\right)\right]p_{r_0}}\right\} \\
&= n_{1k}.
\end{aligned}$$

Similarly, marginal weighted MI produces unbiased estimates of cell counts for $l = 0$, and the method is therefore unbiased for model (3.3)'s parameter estimates.

In conditional weighted MI, the level-l predicted probability of x in the completed data, p_l^{pred}, is given by

$$\begin{aligned}
p_l^{\mathrm{pred}} &= \frac{n_{l+}^{\mathrm{obs}} + n_{l+}^{\mathrm{req}}}{n_{++}} \\
&= \frac{n_{l+}^{\mathrm{obs}} + \sum_{k=0}^{1}\hat{p}_{lk}^{\mathrm{s}}n_{+k}^{\mathrm{mis}}}{n_{++}} \\
&= \frac{\sum_{k=0}^{1} n_{+k}\frac{n_{lk}^{\mathrm{obs}}}{n_{+k}^{\mathrm{obs}}}}{n_{++}}.
\end{aligned} \tag{3.6}$$

Estimated count for cell (l, k) is therefore biased as shown below

$$\begin{aligned}
\hat{n}_{lk}^{\mathrm{c}} &= n_{lk}^{\mathrm{obs}} + n_{lk}^{\mathrm{req}} \\
&= n_{lk}p_{r_l} + n_{+k}^{\mathrm{mis}} \frac{n_{lk}p_{r_l}w_l^{\mathrm{c}}}{\sum_{l=0}^{1} n_{lk}p_{r_l}w_l^{\mathrm{c}}} \\
&= n_{lk}p_{r_l}\left(1 + \frac{n_{+k}^{\mathrm{mis}}w_l^{\mathrm{c}}}{\sum_{l=0}^{1} n_{lk}p_{r_l}w_l^{\mathrm{c}}}\right) \\
&= n_{lk}\hat{b}_{lk}^{\mathrm{c}}.
\end{aligned}$$

Consequently, conditional weighted MI produces biased estimates of model (3.3)'s parameters. In initial two-dimensional simulations performed to verify the calculations when α_0 is fixed to a single value (appendix A.1), bias in conditional weighted MI appears negligible. To confirm

that this bias exists, note that the formulae for marginal and conditional weights are given by

$$w_l^{\mathrm{m}} = \frac{p_l^{\mathrm{pop}} n - p_l^{\mathrm{obs}} n^{\mathrm{obs}}}{n^{\mathrm{mis}} p_l^{\mathrm{obs}}};$$
$$w_l^{\mathrm{c}} = \frac{p_l^{\mathrm{pop}} n - p_l^{\mathrm{pred}} n^{\mathrm{obs}}}{n^{\mathrm{mis}} p_l^{\mathrm{pred}}},$$

respectively, which only differ in p_l^{obs} and p_l^{pred}. Since marginal weighted MI is unbiased under this missingness mechanism, conditional weighted MI is also unbiased if $p_l^{\mathrm{pred}} = p_l^{\mathrm{obs}}$. From (3.6), these two probabilities can be written as

$$\begin{aligned} p_l^{\mathrm{pred}} &= \sum_{k=0}^{1} \frac{n_{+k}}{n_{++}} \cdot \frac{n_{lk}^{\mathrm{obs}}}{n_{+k}^{\mathrm{obs}}} \\ &= \sum_{k=0}^{1} p\left(x = l \mid y = k, r = 1\right) p\left(y = k\right); \\ p_l^{\mathrm{obs}} &= p\left(x = l \mid r = 1\right) \\ &= \sum_{k=0}^{1} p\left(x = l \mid y = k, r = 1\right) p\left(y = k \mid r = 1\right). \end{aligned}$$

Under this missingness mechanism, since missingness in x does not depend on y after conditioning on x, the probability of y conditional on x is therefore the same among the missing and observed x, i.e. $p\left(y = k \mid x = l, r = 1\right) = p\left(y = k \mid x = l\right)$. The marginal probability of y in the observed data can be written as

$$\begin{aligned} p\left(y = k \mid r = 1\right) &= \sum_{l=0}^{1} p\left(y = k \mid x = l, r = 1\right) p\left(x = l \mid r = 1\right) \\ &= \sum_{l=0}^{1} p\left(y = k \mid x = l\right) p\left(x = l \mid r = 1\right) \\ &\neq \sum_{l=0}^{1} p\left(y = k \mid x = l\right) p\left(x = l\right). \end{aligned} \tag{3.7}$$

Since $\sum_{l=0}^{1} p\left(y = k \mid x = l\right) p\left(x = l\right) = p\left(y = k\right)$, (3.7) implies that $p\left(y = k \mid r = 1\right) \neq p\left(y = k\right)$, which means that $p_l^{\mathrm{pred}} \neq p_l^{\mathrm{obs}}$. Hence, bias in conditional weighted MI does exist when missingness in x depends on x.

4. M4 – bias when x is MNAR dependent on x and y
   The cell-wise probability of observing x is dependent on both x and y, $p\left(r_{lk} = 1\right) = p_{r_{lk}}$. Estimated count for cell (l, k) in standard MI is given by

$$\begin{aligned} \hat{n}_{lk}^{\mathrm{s}} &= n_{lk}^{\mathrm{obs}} + n_{lk}^{\mathrm{req}} \\ &= n_{lk}^{\mathrm{obs}} + n_{+k}^{\mathrm{mis}} \frac{n_{lk}^{\mathrm{obs}}}{n_{+k}^{\mathrm{obs}}} \\ &= n_{lk}^{\mathrm{obs}} \frac{n_{+k}}{n_{+k}^{\mathrm{obs}}} \\ &= n_{lk} \left(\frac{p_{r_{lk}} \sum_{l=0}^{1} n_{lk}}{\sum_{l=0}^{1} n_{lk} p_{r_{lk}}} \right) \\ &= n_{lk} \hat{b}_{lk}^{\mathrm{s}}, \end{aligned}$$

   which leads to bias in both parameter estimates of model (3.3).

In marginal weighted MI, the estimated count for cell (l, k) can be written as

$$\begin{aligned}\hat{n}^{\mathrm{m}}_{lk} &= n^{\mathrm{obs}}_{lk} + n^{\mathrm{req}}_{lk} \\ &= n_{lk} p_{r_{lk}} + n^{\mathrm{mis}}_{+k} \frac{n_{lk} p_{r_{lk}} w^{\mathrm{m}}_{l}}{\sum_{l=0}^{1} n_{lk} p_{r_{lk}} w^{\mathrm{m}}_{l}} \\ &= n_{lk} \left(p_{r_{lk}} + \frac{n^{\mathrm{mis}}_{+k} p_{r_{lk}} w^{\mathrm{m}}_{l}}{\sum_{l=0}^{1} n_{lk} p_{r_{lk}} w^{\mathrm{m}}_{l}} \right) \\ &= n_{lk} \hat{b}^{\mathrm{m}}_{lk},\end{aligned}$$

which also results in bias in both parameter estimates of model (3.3).

Similarly in conditional weighted MI, the estimated count for cell (l, k) is

$$\begin{aligned}\hat{n}^{\mathrm{c}}_{lk} &= n^{\mathrm{obs}}_{lk} + n^{\mathrm{req}}_{lk} \\ &= n_{lk} p_{r_{lk}} + n^{\mathrm{mis}}_{+k} \frac{n_{lk} p_{r_{lk}} w^{\mathrm{c}}_{l}}{\sum_{l=0}^{1} n_{lk} p_{r_{lk}} w^{\mathrm{c}}_{l}} \\ &= n_{lk} \left(p_{r_{lk}} + \frac{n^{\mathrm{mis}}_{+k} p_{r_{lk}} w^{\mathrm{c}}_{l}}{\sum_{l=0}^{1} n_{lk} p_{r_{lk}} w^{\mathrm{c}}_{l}} \right) \\ &= n_{lk} \hat{b}^{\mathrm{c}}_{lk}.\end{aligned}$$

Therefore, conditional weighted MI also yields biased parameter estimates of the analysis model (3.3).

3.3.3 *Verification of analytic calculations using simulation*

The analytic calculations presented in section 3.3.2 are verified by simulating a full-data sample with $n = 10\,000$ observations of x and y from the following models

$$\begin{aligned}x &\sim \text{Bernoulli}\left(p^{\mathrm{pop}}_{x} = 0.7\right); \\ \text{logit}\left[p\left(y = 1 \mid x\right)\right] &= \beta_0 + \beta_x x,\end{aligned}$$

where $\beta_0 = \ln(0.5)$ and $\beta_x = \ln(1.5)$. Missing values in x are generated using selection models M1–M4 with a range of values for the selection parameters α (table 3.3).

Bias in $\hat{\beta}_0$ and $\hat{\beta}_x$ is calculated for standard MI and marginal and conditional weighted MI as the difference between $\hat{\beta}_0$ and $\hat{\beta}_x$ and their true values, $\ln(0.5)$ and $\ln(1.5)$, respectively. Parameter estimates are derived analytically by following the calculations in section 3.3.2, and obtained empirically by creating $M = 10$ imputations of missing values in x using the three MI

Table 3.3. Analytic study: values of selection parameters for generating missingness in x used in simulations conducted to verify analytic calculations.

Missingness model	Linear predictor of selection model $\text{logit}\left[p\left[(r = 1 \mid x, y)\right]\right]$	Selection parameter			% missing x
		α_0	α_x	α_y	
M1	α_0	$[-3, 3]$			5–95
M2	$\alpha_0 + \alpha_y y$	$[-3, 3]$		$[-3, 3]$	3–97
M3	$\alpha_0 + \alpha_x x$	$[-3, 3]$	$[-3, 3]$		2–98
M4	$\alpha_0 + \alpha_x x + \alpha_y y$	0.5	$[-3, 3]$	$[-3, 3]$	9–84

* Note: r: response indicator of x.

methods under evaluation. All simulations are performed in Stata 14 [44], where `mi impute logit` [75] is used for standard MI, my command `mi impute wlogit` [73] for marginal and conditional weighted MI, and `mi estimate: logit` [75] for fitting the analysis model to the completed datasets and combining the results using Rubin's rules [20, 21].

Figures 3.1–3.3 present the analytic bias in standard MI, marginal and conditional weighted MI under MAR and MNAR mechanisms with the various values of the selection parameters. Results of the analysis of complete records are also included for comparison.

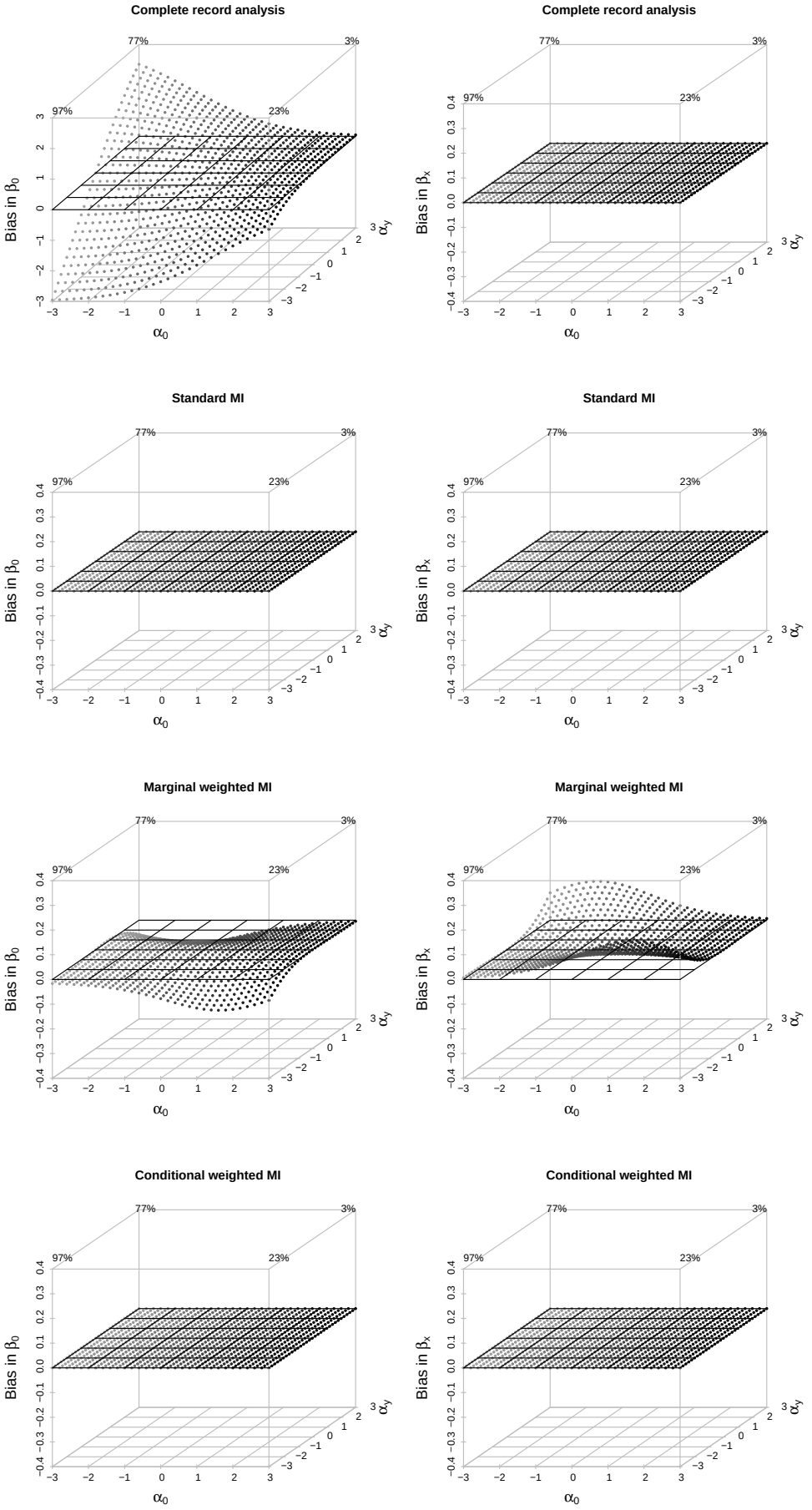
Overall, the empirical results closely match the analytic results (appendix A.1). When x is MCAR (M1), all methods provide unbiased parameter estimates, as expected (appendix A.1).

When x is MAR conditional on y (M2), standard MI and conditional weighted MI are unbiased, while bias is observed for complete record analysis (CRA) in $\hat{\beta}_0$, and for marginal weighted MI in both parameter estimates. This bias is due to the marginal weights not accounting for the association of x and y in the imputation model for x. As a result, marginal weights do not successfully recover the correct distribution of x after MI.

Both parameter estimates are unbiased in marginal weighted MI when x is MNAR dependent on x (M3), while standard MI leads to noticeable bias in the estimate of β_0. Bias in conditional weighted MI is small and occurs for extreme values of the selection parameters. Since missingness in x does not depend on y under M3, CRA is unbiased as the theory suggests.

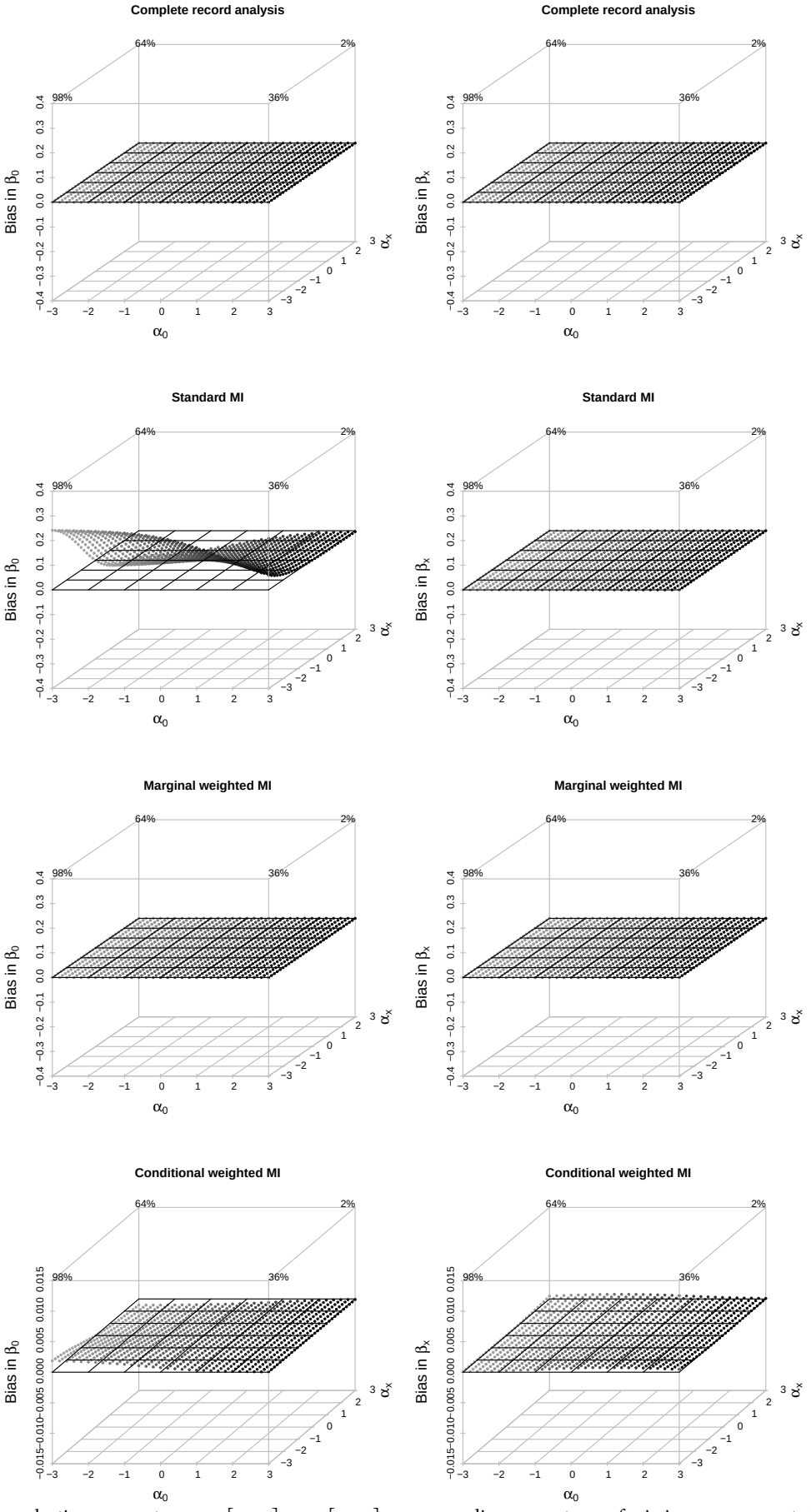
Under the last missingness mechanism where x is MNAR dependent on both x and y (M4), bias in parameter estimates appears to be the smallest in conditional weighted MI. Although bias is present in both standard MI and marginal weighted MI, the magnitude of bias is smaller in marginal weighted MI compared to standard MI. Under this missingness model, conditional weighted MI can be regarded as a hybrid of marginal weighted MI and standard MI. The conditional weights correct for some bias introduced by x in the selection model in a similar manner to the marginal weights under M3; the method also alleviates some residual bias similarly to standard MI under M2.

Figure 3.1. Analytic study: analytic bias when x is MAR conditional on y (M2).



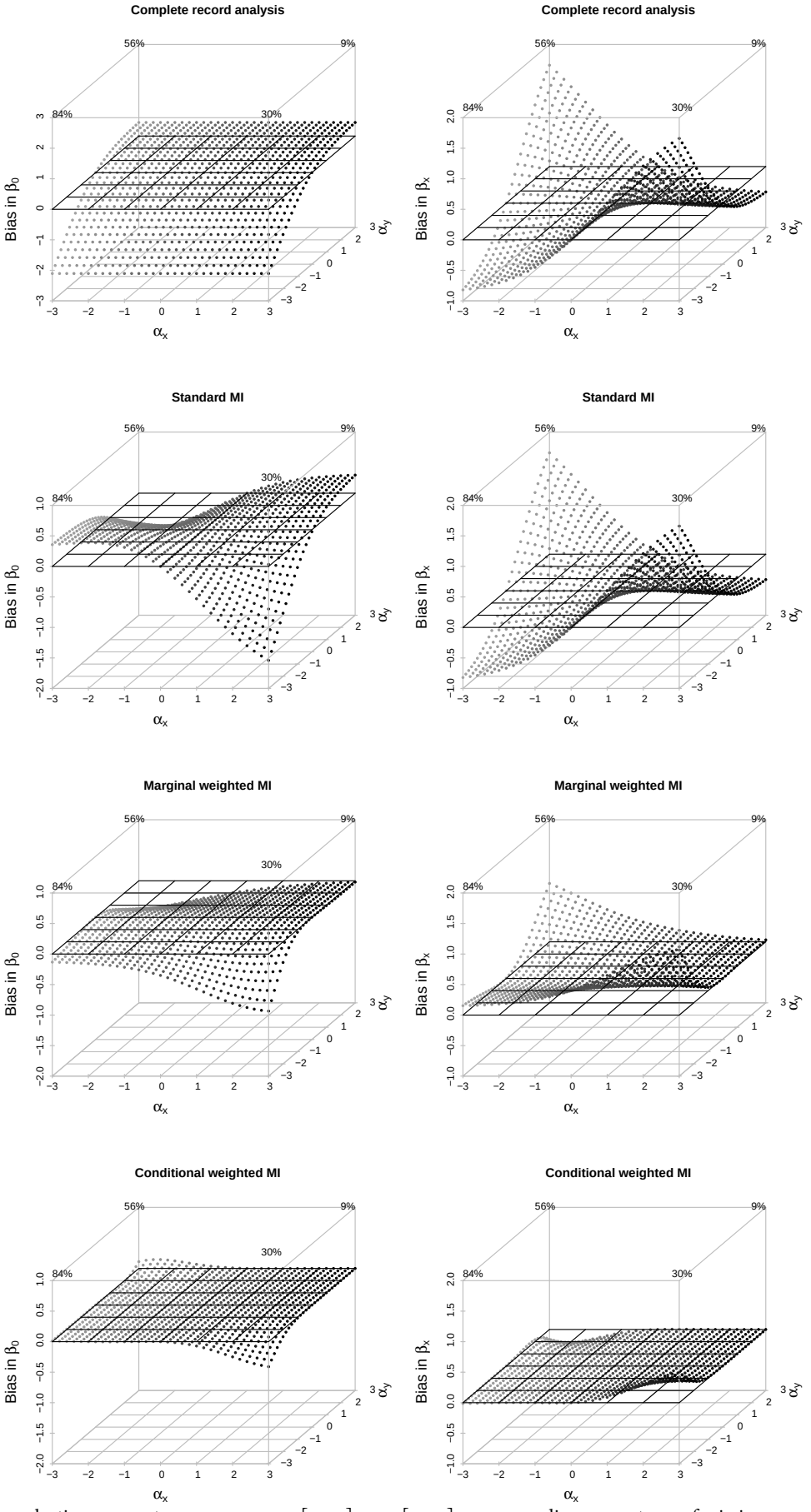
* Note: selection parameters $\alpha_o \in [-3, 3]$, $\alpha_y \in [-3, 3]$; corresponding percentages of missing x are presented for extreme values (± 3) of the α parameters.

Figure 3.2. Analytic study: analytic bias when x is MNAR dependent on x (M3).



* Note: selection parameters $\alpha_o \in [-3, 3]$, $\alpha_x \in [-3, 3]$; corresponding percentages of missing x are presented for extreme values (± 3) of the α parameters.

Figure 3.3. Analytic study: analytic bias when x is MNAR dependent on x and y (M4).



* Note: selection parameters $\alpha_0 = 0.5$, $\alpha_x \in [-3, 3]$, $\alpha_y \in [-3, 3]$; corresponding percentages of missing x are presented for extreme values (± 3) of the α parameters.

This section reports a univariate simulation study conducted to examine other performance measures of weighted MI of a binary covariate, when the fully observed outcome variable is also binary. The aims of this simulation study are to examine finite-sample properties of weighted MI over repeated simulations in terms of bias in parameter estimates, efficiency, and coverage of 95% confidence intervals (CI); and to investigate when assumptions of the method are correct and incorrect.

The order of the simulation studies presented in this thesis is generic and broadly involves the following steps.

1. Generate the 'full data' under the chosen data generating mechanism(s);
2. Make data missing with the chosen missingness mechanism(s);
3. Perform the analysis of interest on complete records (CRA);
4. Impute missing values under the chosen MI method(s) to create M completed datasets;
5. Fit the analysis model to each of the M completed datasets and combine the results using Rubin's rules [20, 21].

Steps 1–5 are then repeated to evaluate frequentist properties of the methods under comparison.

Note that the term 'univariate' is used here to refer to the nature of missing data in the analysis, where values are missing in a single incomplete covariate. Multivariate simulation studies are presented later in this chapter; the term 'multivariate' refers to settings where missing values occur in more than one covariate. These are consistent with the use of terminologies for describing missingness patterns in section 2.2.

3.4.1 *Method*

Similar to the analytic study presented in section 3.3, the analysis model in this simulation study is a logistic regression of a fully observed binary outcome y on an incomplete binary covariate x. As before, marginal and conditional weighted MI are compared to standard MI under different missingness mechanisms. The data generating mechanism and analysis procedures are as follows.

1. Simulate $n = 5\,000$ complete values of the binary $\{0, 1\}$ covariate x and binary $\{0, 1\}$ outcome y from the models

$$\begin{aligned} x &\sim \text{Bernoulli}\left(p_x^{\text{pop}} = 0.7\right); \\ \text{logit}\left[p\left(y = 1 \mid x\right)\right] &= \beta_0 + \beta_x x, \end{aligned} \tag{3.8}$$

   where β_0 and β_x are arbitrarily set to $\ln(0.5)$ and $\ln(1.5)$, respectively. The same values of the β parameters are used throughout to make bias comparable across all simulation settings. This sample size is chosen to minimise the issue of small-sample bias associated with the logistic regression [76];
2. Simulate a binary indicator of response r of x from each of the selection models M1–M4 (table 3.2). Values of 1.5 and -1.5 are chosen for α_y and α_x in M2 and M3, respectively, to reflect strong odds ratios (OR) of observing x (OR = 4.48 and 0.22, respectively). For M4, $\alpha_y = 1.5$ and $\alpha_x = -1.5$ are chosen as, according to the results depicted in figure 3.3, bias in the three MI methods under evaluation is likely to be apparent with these coefficients predicting

missingness in x. For all selection models, α_0 is altered to achieve approximately 45% missing x. For M1, α_0 is calculated directly as $\ln\left(\frac{0.55}{0.45}\right)$; for M2–M4, $\alpha_0 = -0.2; 1.35;$ and 0.75 appear to work well;

3. For $i = 1, \ldots, 5\,000$, set x_i to missing if $r_i = 0$;
4. Impute missing values in x $M = 50$ times using standard MI, marginal and conditional weighted MI in turn;
5. For each MI method, fit the analysis model (3.8) to each completed dataset and combine the results using Rubin's rules [20, 21].

Steps 1–5 are repeated $S = 1\,000$ times under each of the four missingness models M1–M4, so the same set of simulated independent datasets is used to compare the three MI methods under the same missingness scenario, but a different set of datasets is generated for each missingness scenario [77]. The parameters of interest are β_0 and β_x, although in practice β_x is usually of more interest. Bias, efficiency of $\hat{\beta}_0$ and $\hat{\beta}_x$ in terms of the empirical standard errors, and coverage of 95% CIs are calculated over 1 000 repetitions for each combination of simulation settings [78], with analyses of the full data (i.e. before any data are set to missing) and complete records also provided for reference. All simulations are performed in Stata 14 [44] and simulated datasets are analysed using the community-contributed command `simsum` [78].

3.4.2 *Performance measures*

Assume that the true parameter of interest is β and that the sth simulated dataset $(s = 1, \ldots, S)$ yields a point estimate $\hat{\beta}_s$ with standard error (SE) SE_s. Define the following quantities

$$\overline{\hat{\beta}} = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_s;$$

$$\widehat{\mathrm{Var}}\left(\hat{\beta}\right) = \frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\beta}_s - \overline{\beta}\right)^2;$$

Three performance measures over repeated simulations are summarised below, with Monte-Carlo standard error (MCSE) defined as the standard deviation of an estimated quantity over repeated simulations [78].

1. Bias in point estimate is estimated as

$$\mathrm{Bias} = \overline{\hat{\beta}} - \beta;$$

$$\mathrm{MCSE} = \sqrt{\frac{\widehat{\mathrm{Var}}\left(\hat{\beta}\right)}{S}}.$$

2a. The empirical standard error is estimated as the standard deviation of $\hat{\beta}$ over S repetitions

$$\text{Empirical SE} = \sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}\right)};$$

$$\mathrm{MCSE} = \sqrt{\frac{\widehat{\mathrm{Var}}\left(\hat{\beta}\right)}{2\left(S-1\right)}}.$$

2b. The average model standard error is defined as

$$\overline{\mathrm{SE}} = \frac{1}{S}\sum_{s=1}^{S}\mathrm{SE}_s;$$

$$\text{MCSE} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\text{SE}_s - \overline{\text{SE}}\right)^2}.$$

3. Coverage of 95% CIs is defined as the percentage of times the 95% CIs of $\hat{\beta}_s$ contain the true value of β

$$\text{Coverage } C = \frac{1}{S}\sum_{s=1}^{S} I\left[\mid \hat{\beta}_s - \beta \mid < z_{0.025}\text{SE}_s\right];$$

$$\text{MCSE} = \sqrt{\frac{C\,(1-C)}{S}},$$

where $I\,[\,]$ denotes the indicator function taking values 1 if the statement inside the brackets is true and 0 otherwise, and $z_{0.025}$ is the 0.025 quantile of the standard normal distribution.

3.4.3 Results

Results of the univariate simulation study are summarised graphically in figures 3.4–3.6. The full data and CRA both give the results that the theory predicts. Analysis of the full data is always unbiased with coverage close to the 95% level and the smallest standard errors of all methods. CRA is unbiased under M1 and M3 as expected [78], but bias is observed under the other two missingness mechanisms. Coverage is correspondingly low when bias is present, and efficiency is lower than that in the full data.

Under M1, when x is MCAR, all methods appear unbiased with the empirical standard errors slightly larger than the average model standard errors and correct coverage. This is as expected.

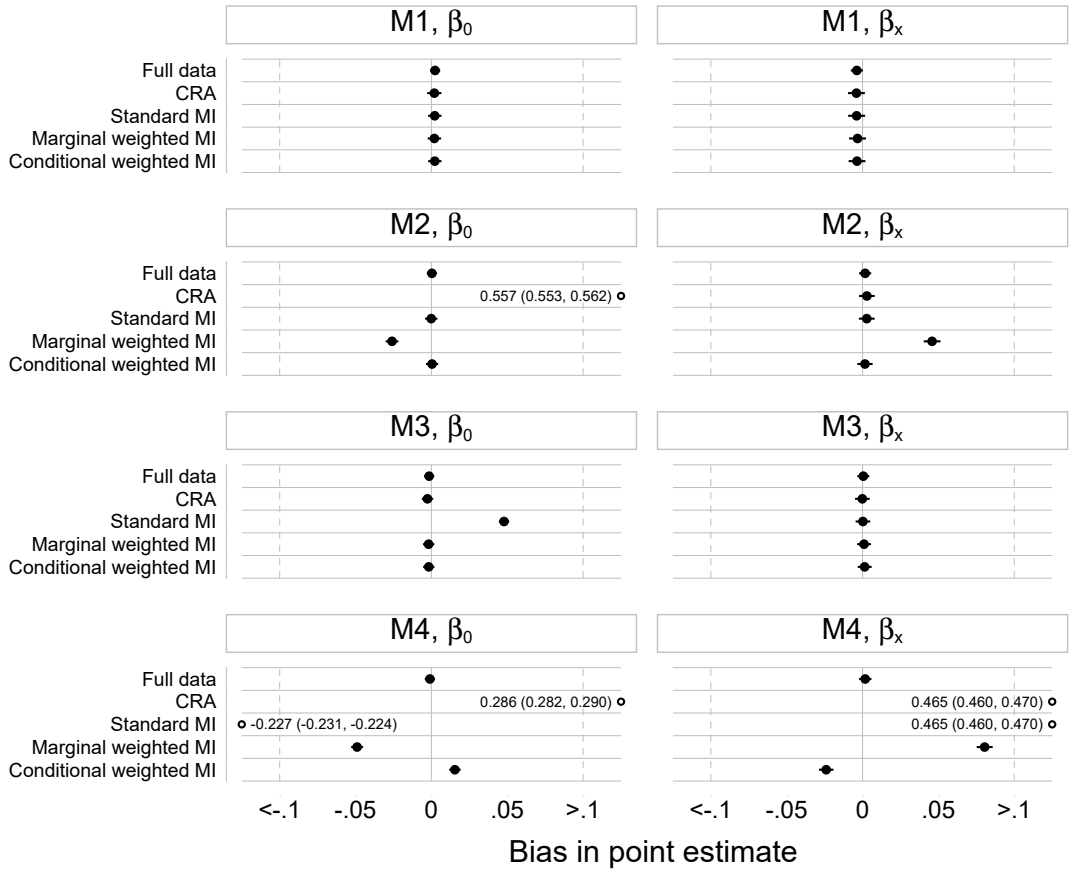
Under M2, when x is MAR conditional on y, only standard MI and conditional weighted MI are unbiased for both parameter estimates. CRA is severely biased in the estimate of β_0 and the corresponding coverage of 95% CIs falls to 0. However, the method provides an unbiased estimate of β_x with correct coverage. This result is specific to this simulation set-up, where the probability of being a complete record depends on the outcome, and the analysis model is a logistic regression. This mimics case-control sampling, where the log odds of the logistic regression are biased in case-control studies but the log odds ratio is not [39, 79]. The covariate–outcome association can therefore be estimated consistently among the complete records. The average model standard errors are similar in standard MI and conditional weighted MI, and they are comparable to their empirical counterparts. This results in correct coverage of 95% CIs. Bias is seen in both parameter estimates in marginal weighted MI, which may explain the discrepancy between the empirical and average model standard errors, as well as the drop in coverage.

Under M3, when x is MNAR dependent on x, standard MI is biased in the estimate of β_0 but provides an unbiased estimate of β_x, which agrees with findings of the analytic study. Since missingness in the covariate does not depend on the outcome, conditional on the covariate, CRA also yields unbiased parameter estimates. Generally, in logistic regression with an incomplete covariate x, when the missingness mechanism is such that both standard MI and CRA are unbiased, standard MI tends not to be more efficient than CRA in estimating β_x [39]. This is because without auxiliary variables in the imputation model, standard MI does not carry any extra information on the odds ratio compared to CRA. This is seen in the simulation results for β_x under models M1–M3. Given that CRA, standard MI, and marginal weighted MI are unbiased

for $\hat{\beta}_x$ under M3, there is a small gain in efficiency in the estimate of β_x in marginal weighted MI, as the empirical standard error is slightly smaller in marginal weighted MI. The efficiency gain in marginal weighted MI compared to CRA is expected to be more apparent in multivariate missing data settings, particularly for incomplete covariates with lower levels of missing data [39]. Although bias in conditional weighted MI is confirmed in the analytic study, this bias appears to be minuscule over repeated simulations, and results in terms of the standard errors and coverage are generally similar to that in marginal weighted MI. When the methods under evaluation are unbiased, their corresponding coverage of 95% CIs generally attains the nominal level.

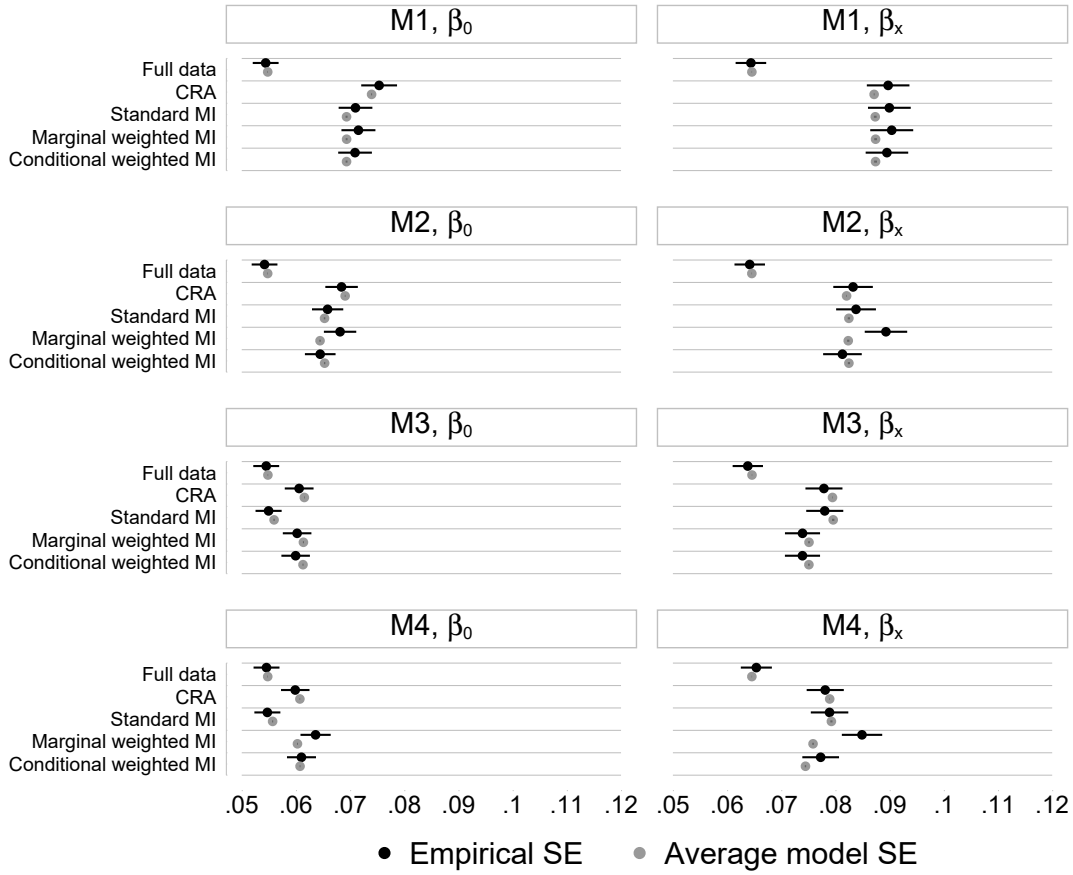
Under M4, when x is MNAR dependent on x and y, standard MI and CRA are again markedly biased in both parameter estimates, leading to coverage close or equal to 0. Conditional weighted MI is the least biased method, whose coverage remains relatively high at just below the 95% level. Although present, bias in marginal weighted MI is less pronounced compared to standard MI and CRA. Marginal weighted MI corrects bias introduced when missingness in x only depends on the values of x, but fails to address bias introduced by the presence of y in the missingness model for x. Coverage therefore falls to below the 90% mark.

Figure 3.4. Univariate simulation study: bias in point estimates under different missingness mechanisms for x.



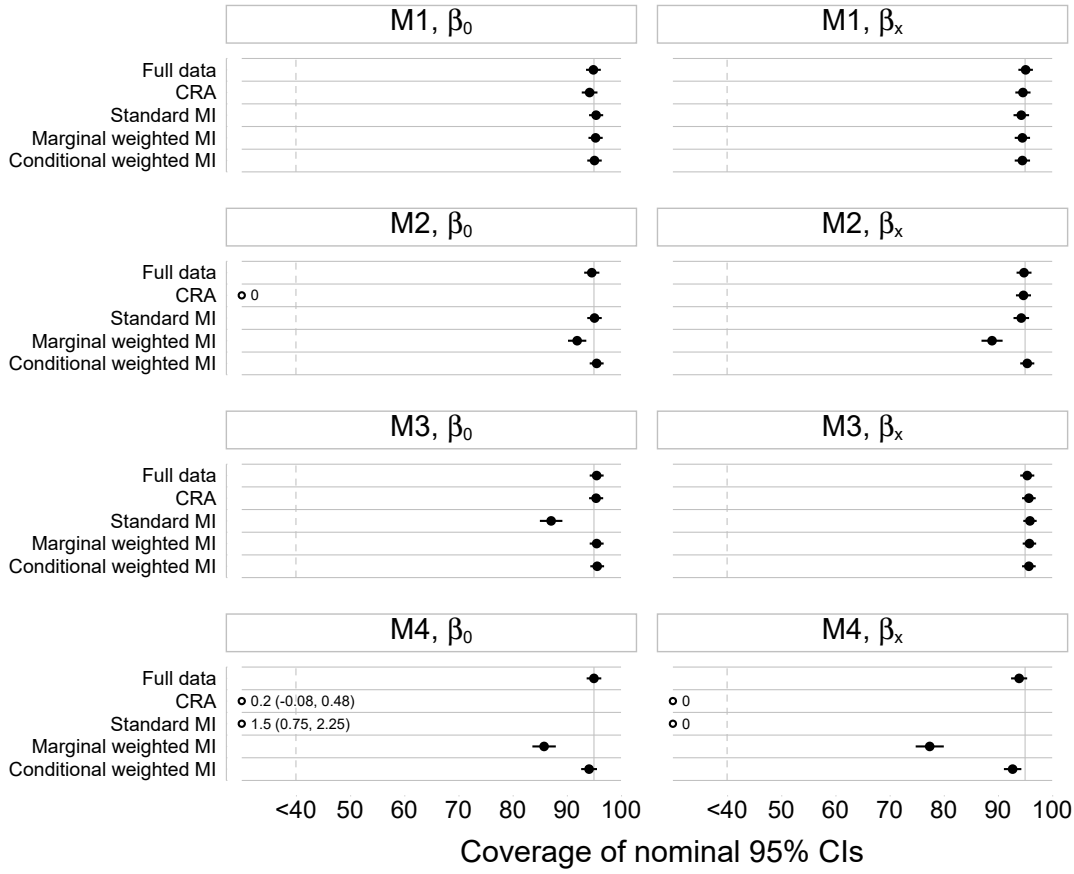
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); $\beta_0 = -0.693$, $\beta_x = 0.405$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure 3.5. Univariate simulation study: empirical and average model standard errors under different missingness mechanisms for x.



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 3.6. Univariate simulation study: coverage of nominal 95% confidence intervals under different missingness mechanisms for x.



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

3.5 Extended univariate simulation study: when there is uncertainty in estimating the population distribution

So far, the population distribution of the incomplete covariate used to derive the weights in weighted MI is assumed to be obtained from a population census or equivalent. In other words, it is assumed that there is no uncertainty associated with estimating the reference distribution, and hence, the weights. In weighted MI, the uncertainty in weights should be ignored when the population distribution of the incomplete variable is assumed to be 'known', unless the reference population is not a census or equivalent. Since MI is a Bayesian procedure in which all sources of uncertainty are modelled, this explains why, if there is no uncertainty about the population distribution of the incomplete variable, weights can be calculated once and used across all imputations.

When there is uncertainty in estimating the population distribution of an incomplete binary/categorical covariate, it raises a question of how this uncertainty should be incorporated in the imputation process. A natural approach for dealing with this extra source of uncertainty would be to draw values of the population proportions from their distribution and calculate the weights using these draws, so that this uncertainty is reflected in the MI variance estimation.

This additional step is expected to have an effect on the between-imputation variance of Rubin's variance estimator.

An extension of the univariate simulation study presented in section 3.4 is discussed next. This extended simulation study explores the setting where the reference distribution is not 'known' and is estimated from an external dataset, e.g. when the reference distribution is assumed to be estimated in an external population survey instead of a census.

3.5.1 *Method*

This extended univariate simulation study of a fully observed binary outcome y and a partially observed binary covariate x follows the same method described in section 3.4.1, except that two variations of the population proportions of x are evaluated in the imputation step of marginal and conditional weighted MI. The reference distribution is assumed to either come from a census or equivalent (case 1), or be estimated in an external dataset of larger size (case 2) or smaller size (case 3) than the study sample.

Suppose that in an external dataset of size n^{ex} which comes from the same population as the study sample, the sample proportion $\hat{p}_x^{\text{pop}}$ provides an unbiased estimate of the population proportion p_x^{pop}. Assuming that the sampling distribution of the sample proportions is approximately normal, its standard error is given by

$$\text{SE}\left(\hat{p}_x^{\text{pop}}\right) = \sqrt{\frac{\hat{p}_x^{\text{pop}}\left(1-\hat{p}_x^{\text{pop}}\right)}{n^{\text{ex}}}}.$$

The data generating mechanism and analysis procedures are as follows.

1. For cases 2 and 3, the following two steps are performed to incorporate the sampling behaviour of $\hat{p}_x^{\text{pop}}$, which is estimated in an external dataset of size n^{ex}, into the data generating mechanism in repeated simulations.
   a. Simulate $n^{\text{ex}} = 10\,000$ (case 2) or $1\,000$ (case 3) complete values of the binary $\{0,1\}$ covariate x from the model
   $$x \sim \text{Bernoulli}\left(p_x^{\text{pop}} = 0.7\right);$$
   b. Obtain the sample proportion, $\hat{p}_x^{\text{pop}}$, of x, which is an unbiased estimate of the population proportion, p_x^{pop};
2. Simulate $n = 5\,000$ complete values of the binary $\{0,1\}$ covariate x and binary $\{0,1\}$ outcome y from the models
   $$\begin{aligned} x &\sim \text{Bernoulli}\left(p_x^{\text{pop}} = 0.7\right); \\ \text{logit}\left[p\left(y=1 \mid x\right)\right] &= \beta_0 + \beta_x x, \end{aligned} \tag{3.9}$$
   where β_0 and β_x are arbitrarily set to $\ln(0.5)$ and $\ln(1.5)$, respectively. The same values of the β coefficients are used throughout to make bias comparable across all simulation settings;
3. Simulate a binary indicator of response r of x from each of the selection models M1–M4 (table 3.2). Values of 1.5 and -1.5 are chosen for α_y and α_x in M2 and M3, respectively. For M4, $\alpha_y = 1.5$ and $\alpha_x = -1.5$ are used. In all selection models, α_0 is altered to achieve approximately 45% missing x. For M1, α_0 is calculated directly as $\ln\left(\frac{0.55}{0.45}\right)$; for M2–M4, $\alpha_0 = -0.2; 1.35;$ and 0.75 are used;

4. For $i = 1, \ldots, 5\,000$, set x_i to missing if $r_i = 0$;
5. Impute missing values in x $M = 50$ times using marginal and conditional weighted MI in turn, by following the below steps for each of the imputations in cases 2 and 3.
   a. Draw a value $\dot{p}_x^{\text{pop}}$ from the normal approximation $\text{N}\left(\hat{p}_x^{\text{pop}}, \frac{\hat{p}_x^{\text{pop}}\left(1-\hat{p}_x^{\text{pop}}\right)}{n^{\text{ex}}}\right)$, with values of $n^{\text{ex}} = 10\,000$ (case 2) and $1\,000$ (case 3). This is done by first taking a draw $\dot{u}$ from the standard normal distribution, $u \sim \text{N}(0, 1)$, followed by calculating the draw $\dot{p}_x^{\text{pop}} = \hat{p}_x^{\text{pop}} + \dot{u}\sqrt{\frac{\hat{p}_x^{\text{pop}}\left(1-\hat{p}_x^{\text{pop}}\right)}{n^{\text{ex}}}}$;
   b. Calculate the marginal and conditional weights with $\dot{p}_x^{\text{pop}}$ as the reference proportion;
   c. Perform marginal and conditional weighted MI using these weights;
6. For each MI method, fit the analysis model (3.9) to each completed dataset and combine the results using Rubin's rules [20, 21].

Step 5 is designed to mimic the full Bayesian sampling process, which is always the aim in MI. Again, steps 1–6 are repeated $S = 1\,000$ times under each of the four missingness models M1–M4, so the same set of simulated independent datasets is used to compare the three MI methods under the same missingness scenario, but a different set of datasets is generated for each missingness scenario [77]. The parameters of interest are β_0 and β_x. Bias in $\hat{\beta}_0$ and $\hat{\beta}_x$, efficiency in terms of the empirical standard errors, and coverage of 95% CIs are calculated over 1000 repetitions for each combination of simulation settings [78], with analyses of the full data and complete records also provided for comparison.

All simulations are performed in Stata 14 [44]. For cases 2 and 3 in step 5, marginal and conditional weighted MI are performed using an adaptation of my command `mi impute wlogit` [73].

3.5.2 *Results*

Figures 3.7–3.9 summarise the results of the extended univariate simulation study for marginal and conditional weighted MI under the various missingness mechanisms for x. Results for both weighted MI methods in case 1 are discussed in section 3.4.3.

Under M3, when missingness in x depends only on x, the imputation process encounters negative weights in weighted MI in 149 (15%) repetitions when $n^{\text{ex}} = 1\,000$. Some of the incomplete datasets with negative weights are recreated from the corresponding states of the random number generator, and values of the negative weights are very close to 0. With the chosen coefficients for generating missingness in x, the majority of missing values in x occur for $x = 1$ ($\alpha_x = -1.5$). This missingness mechanism, together with the increased variation in the estimated proportions of x when $n^{\text{ex}} = 1\,000$, can lead to negative weights where $\dot{p}_x^{\text{pop}} n < p_x^{\text{obs}} n^{\text{obs}}$. Similarly, under M4 when missingness in x depends on both x and y, negative weights are encountered in weighted MI in 161 (16%) repetitions, also when $n^{\text{ex}} = 1\,000$. The negative weights w_0 are set to 0.00001; the log odds of the imputation model are therefore increased by a large factor $\ln\left(\frac{w_1}{w_0}\right)$, which ensures that the predicted probability of $x = 1$ in the imputed data is very close to 1.

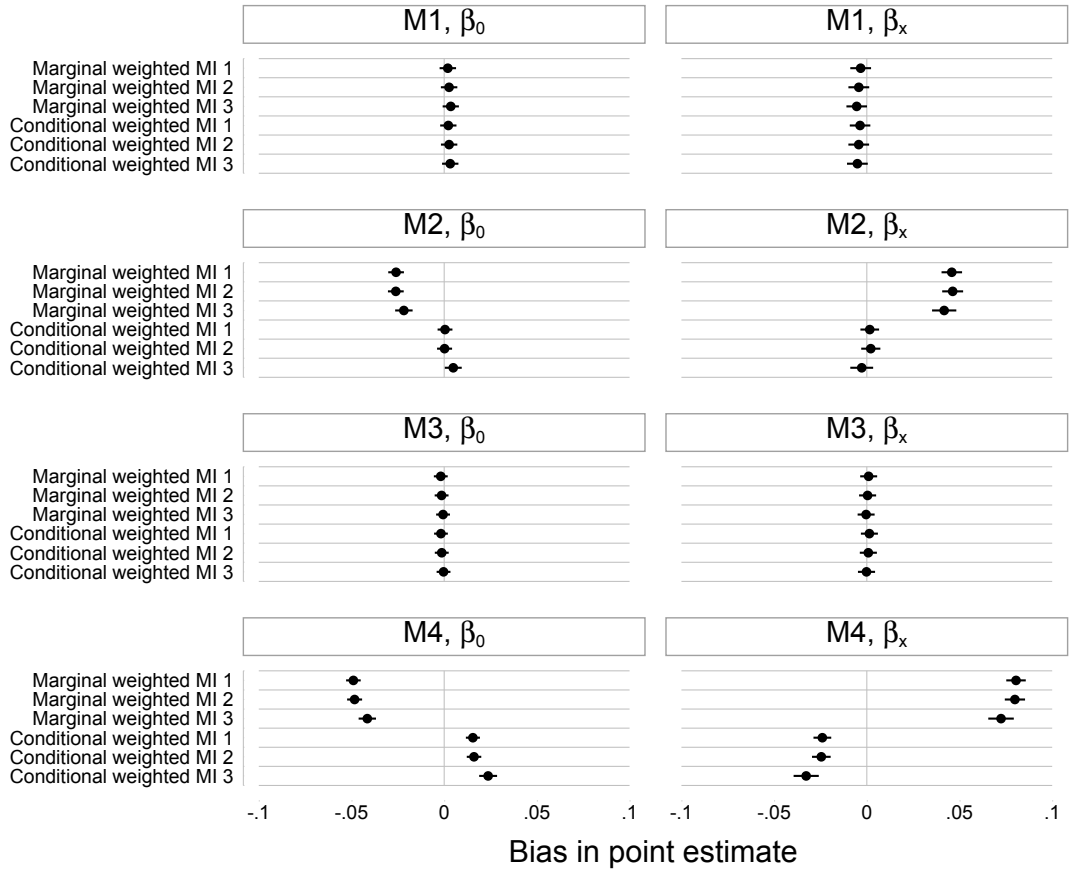
Bias in point estimates is similar when p_x^{pop} is invariant or estimated in a large external dataset (cases 1 and 2). Bias slightly increases, particularly under M2 and M4, when p_x^{pop} is estimated in a small external dataset with higher variance (case 3, figure 3.7).

Empirical standard errors and average model standard errors are comparable and remain

stable for both weighted MI methods across the three cases under M1 and M3. Under M2 and M4, the discrepancy between the empirical and average model standard errors in marginal weighted MI, which is thought to be caused by bias in the point estimates, is seen in all three cases. Empirical and average model standard errors are similar in conditional weighted MI in all cases. However, when there is increased uncertainty in estimating the population proportions of x (case 3 compared to case 1), there is also a marked increase in the empirical and average model standard errors in both marginal and conditional weighted MI. This extra uncertainty is reflected in the variation of the point estimates across the simulation repetitions according to how the simulation is set up, and is also acknowledged in the between-imputation variance component of Rubin's variance estimator (table 3.4).

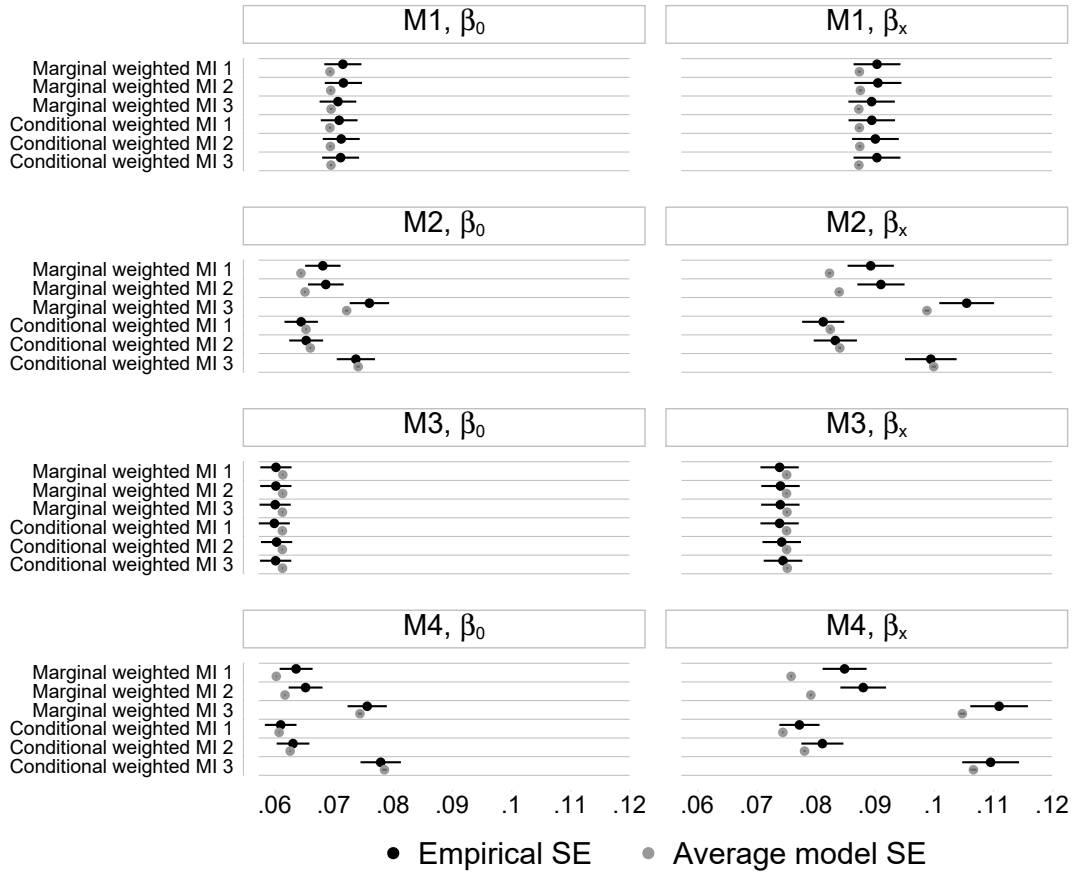
In line with the results seen for the standard errors, coverage attains the nominal level for both weighted MI methods under M1 and M3. Under M2, bias in marginal weighted MI leads to a slight under-coverage in the first two cases. In case 3 and under M2, due to the increase in standard errors in marginal weighted MI, coverage is slightly improved and is closer to the 95% level. There is also an increase in coverage in both marginal and conditional weighted MI under M4. This increase corresponds to the larger standard errors, which are associated with the higher uncertainty in estimating p_x^{pop}.

Figure 3.7. Extended univariate simulation study: bias in point estimates under different missingness mechanisms for x; the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).



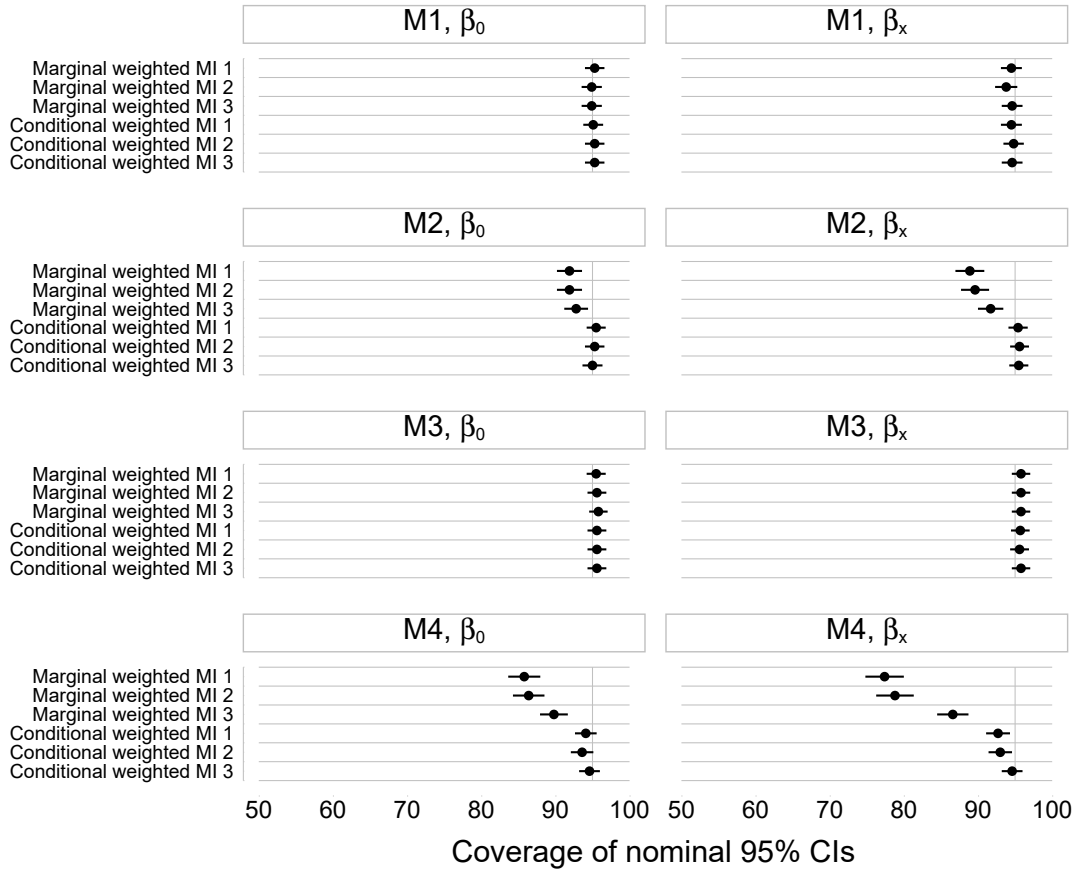
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); $\beta_0 = -0.693$, $\beta_x = 0.405$; error bars: ±1.96× Monte Carlo standard errors.

Figure 3.8. Extended univariate simulation study: empirical and average model standard errors under different missingness mechanisms for x; the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 3.9. Extended univariate simulation study: coverage of nominal 95% confidence intervals under different missingness mechanisms for x; the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Table 3.4. Extended univariate simulation study: variance information about the β parameter estimates in marginal weighted MI in one simulation repetition, when x is MNAR dependent on x and y (M4); the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).

a. Case 1

	$\widehat{W}$	$\widehat{B}$	Total	RVI	FMI	RE
$\hat{\beta}_0$	0.00284	0.00057	0.0034	0.2052	0.1712	0.9966
MC error	<0.0001	0.00013	0.0001	0.0468	0.0330	0.0007
$\hat{\beta}_x$	0.00404	0.00124	0.0053	0.3147	0.2411	0.9952
MC error	<0.0001	0.00028	0.0003	0.0714	0.0426	0.0008

b. Case 2

	$\widehat{W}$	$\widehat{B}$	Total	RVI	FMI	RE
$\hat{\beta}_0$	0.00284	0.00084	0.0037	0.3009	0.2330	0.9954
MC error	<0.0001	0.00015	0.0002	0.0538	0.0325	0.0007
$\hat{\beta}_x$	0.00403	0.00192	0.0060	0.4838	0.3290	0.9935
MC error	<0.0001	0.00035	0.0004	0.0882	0.0412	0.0008

c. Case 3

	$\widehat{W}$	$\widehat{B}$	Total	RVI	FMI	RE
$\hat{\beta}_0$	0.00298	0.00305	0.0061	1.0436	0.5158	0.9898
MC error	<0.0001	0.00061	0.0006	0.2082	0.0516	0.0010
$\hat{\beta}_x$	0.00415	0.00696	0.0112	1.7099	0.6368	0.9874
MC error	<0.0001	0.00138	0.0014	0.3387	0.0476	0.0009

Note: $\widehat{W}$: within-imputation variance; $\widehat{B}$: between-imputation variance; Total: total variance; RVI: relative increase in variance; FMI: fraction of missing information; RE: relative efficiency; MC error: Monte Carlo error.

3.5.3 *Univariate simulation studies: conclusion and remarks*

Results of the univariate simulation studies presented in sections 3.4 and 3.5 confirm the findings of the analytic study (section 3.3) in terms of bias in the analysis model's parameter estimates. In a 2 × 2 contingency table setting where the outcome variable is fully observed and the incomplete covariate is MAR conditional on the outcome, standard MI and conditional weighted MI are equivalent and are the preferred methods.

When values of the covariate are MNAR, marginal weighted MI and CRA are preferred to standard MI and conditional weighted MI when missingness does not depend on the outcome. Given that bias in conditional weighted MI under this missingness mechanism is confirmed in the analytic calculations, simulation results show that this bias appears negligible, and marginal and conditional weighted MI produce more or less the same results. CRA is an unbiased method for handling missing data under this missingness mechanism and its standard errors are relatively similar to that in marginal weighted MI. Nevertheless, the latter can potentially lead to a gain in efficiency when there is more than one incomplete covariate in a higher-order contingency table.

When missingness in the covariate depends on both the values of the covariate and the

outcome, none of the weighted MI methods are unbiased, but they both outperform CRA and standard MI in terms of bias and coverage. Although it is not clear from the analytic calculations which of the two weighted MI methods yields smaller bias in expectation under this missingness mechanism, it appears in the simulation studies that bias is smaller in conditional weighted MI. This confirms that the effects of covariates in the imputation model need to be taken into account when deriving the weights. Nevertheless, the fact that bias is still observed in conditional weighted MI under this missingness mechanism indicates that the proposed conditional weights might not be an entirely optimal approach to account for the effects of covariates in the imputation model.

Further simulations (seciton 3.5) show that when the population proportions of the covariate are estimated in a small external dataset with a higher level of uncertainty, there is an increase in the empirical and average model standard errors in both marginal and conditional weighted MI, particularly when missingness depends on the outcome (missingness models M2 and M4). This leads to an increase in the coverage of 95% CIs. The increase in the average model standard errors is due to an increase in the between-imputation variance in Rubin's variance estimator. Results from this extended simulation study suggest that the extra uncertainty arising from drawing the population proportions from their distribution and calculating the weights is reflected in Rubin's variance estimator. However, further investigations are required to understand why this increase in uncertainty is more noticeable when missingness depends on the outcome (M2 and M4).

For application in case studies presented in sections 6.4 and 6.5, it is worth highlighting that findings in the analytic and univariate simulation studies can be generalised to the case where the incomplete covariate is a categorical variable with more than two levels. Suppose that in a two-way contingency table, the fully observed outcome variable y is binary taking values $k = 0$ or 1 as before, but the partially observed covariate x is now an L-level categorical variable. The analysis model is still a logistic regression of y on x, but the imputation model for the L-level variable x is now a multinomial logistic regression of x conditional on the binary outcome y. Setting the highest level of x, e.g. the Lth level, as the base level to define the model, each level of x can be modelled with a logistic regression

$$\log\left(\frac{p_{x_{lk}}}{p_{x_{Lk}}}\right) = \alpha_{0_l} + \alpha_{y_l} I\left[y = k\right], \tag{3.10}$$

for $l = 1, \ldots, L-1$ and $k = 0, 1$.

This model is analogous to the logistic regression imputation model used previously when the incomplete covariate x was binary, except that the probability distribution of the covariate is now multinomial instead of binomial, and there are $L-1$ equations instead of one. These equations contrast each of the $l = 1, \ldots, L-1$ levels with level L, whereas the single logistic regression model is a contrast between only two levels, i.e. the multinomial logistic regression model reduces to the usual logistic regression model considered previously if $L = 2$. The covariate–outcome association in the imputation model can still be expressed in terms of the odds ratios when each level of the incomplete covariate is contrasted to the base level. This supports the generalisability of results in the analytic and univariate simulation studies to the case of an incomplete categorical covariate. Further simulations can be performed to confirm this generalisability. The above remark permits the theoretical results to be extended to the situation of an incomplete categorical variable, and thus strongly suggests that simulation results would give similar findings.

Until now, the development and evaluation of the marginal and conditional weighted MI methods are restricted to the case of a single partially observed covariate. This section considers an extension of weighted MI in the multivariate imputation by chained equations (MICE) approach [23] (section 2.4.3) for imputing missing values in more than one covariate. In particular, the proposed univariate marginal and conditional weighted MI methods can be embedded into the chained equations to impute covariates whose population marginal distributions are available externally, while the standard (unweighted) MI method can be used for the imputation of other covariates. When there are several variables to be imputed using their reference distributions, information from more than one external data source can be utilised in weighted MI of these variables.

3.6.1 Imputation procedure

The MICE algorithm is available in Stata from version 12 onward via the command `mi impute chained`. However, `mi impute chained` only allows for the specification of a global set of probability weights `pweight`. This set of weights is applied to every univariate conditional model in the algorithm [75]. Suppose it is of interest to examine risk factors of diabetes, and missing data occur in both covariates of interest, namely ethnicity and smoking status. The population distribution of ethnicity is available from the census, but the population distribution of smoking is not. Hence, it is not intuitive to use weights calculated from the population distribution of ethnicity to impute missing smoking status.

Based on the Stata community-contributed command `uvis` [74], I have written Stata commands to perform multivariate imputation by chained equations in the same manner as `mi impute chained`, but allowing for a marginal or conditional weight specification option in each conditional model separately, as illustrated below for the incomplete variable z_j, $j = 1 \ldots, q$ at iteration $t + 1$.

$$
\begin{aligned}
&z_1^{(t+1)} \mid z_2^{(t)}, z_3^{(t)}, \ldots, z_q^{(t)} \quad \longleftarrow \text{weights} = w_{z_1};\\
&z_2^{(t+1)} \mid z_1^{(t+1)}, z_3^{(t)}, \ldots, z_q^{(t)} \quad \longleftarrow \text{weights} = 1;\\
&\vdots\\
&z_q^{(t+1)} \mid z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{q-1}^{(t+1)} \quad \longleftarrow \text{weights} = 1.
\end{aligned}
$$

In the above example, users can specify marginal or conditional weights in the conditional model for ethnicity, given the complete outcome and observed and previously imputed values of smoking status from the last iteration. Subsequently, these weights can be removed from the conditional model for imputing smoking status.

In MICE with marginal weighted conditional models (referred to as *marginal weighted MICE*), weights for z_j are calculated once using the observed and population distributions of z_j and applied in all imputations. In MICE with conditional weighted conditional models (referred to as *conditional weighted MICE*), weights for z_j are calculated at the beginning of the iterative process using the observed and filled-in values of the other variables z_{-j}, and then re-calculated at each iteration using the observed and previously imputed values of z_{-j} from the last iteration. The

calculation of the conditional weights is performed in each imputation, since missing values in the incomplete variables are filled in randomly using observed values at the beginning of the algorithm and then imputed values are updated throughout the iterative process.

3.6.2 *Method*

This section discusses a single multivariate simulation study that extends the previous 2×2 contingency table setting into a three-way contingency table, considering different assumptions of the missingness mechanism for the partially observed covariates.

The analysis model in this simulation study is a logistic regression of a fully observed binary outcome y on two incomplete binary covariates x and z. As before, marginal and conditional weighted MICE are compared to standard MICE under the various missingness mechanisms M1–M5 for x and z (table 3.5). These missingness mechanisms are not the exhaustive set of all the possible mechanisms for missingness in x and z, but they represent the mechanisms that are relatively straightforward to describe and interpret.

The data generating mechanism and analysis procedures are as follows.

1. Simulate $n = 100\,000$ complete observations of the binary covariates x and z from the models

$$z \sim \text{Bernoulli}\left(p_z^{\text{pop}} = 0.7\right);$$
$$\text{logit}\left[p\left(x \mid z\right)\right] = \beta_{x0} + \beta_{xz} z,$$

   where $\beta_{x0} = -0.5$ and $\beta_{xz} = 0.5$ are used for a moderate association between x and z (odds ratio = 1.6). The sample size is chosen to demonstrate large-sample bias in the point estimates if present. The true population proportion of x, p_x^{pop}, is given by

$$\begin{aligned} p_x^{\text{pop}} &= p\left(x = 1 \mid z = 0\right) p\left(z = 0\right) + p\left(x = 1 \mid z = 1\right) p\left(z = 1\right) \\ &= \text{expit}\left(-0.5\right) \cdot 0.3 + \text{expit}\left(0\right) \cdot 0.7; \end{aligned}$$

2. Simulate complete data of the binary outcome y from the logistic regression model

$$\text{logit}\left[p\left(y \mid x, z\right)\right] = \beta_{y0} + \beta_{yx} x + \beta_{yz} z, \tag{3.11}$$

   where $\beta_{y0}, \beta_{yx}, \beta_{yz}$ are arbitrarily set to 0.5, −1, and 1, respectively. The same values of the β_y

Table 3.5. Single multivariate simulation study: variables associated with missingness in x and z, corresponding selection parameters, and percentages of observed data in x and z.

Variables associated with missingness		Selection parameters						% observed data				Label
x	z	α_{x0}	α_{xx}	α_{xy}	α_{z0}	α_{zz}	α_{zy}	(x,z)	$(x,\cdot)$	$(\cdot,z)$	$(\cdot,\cdot)$	
x	y	0.5	1.5		−0.15		1.5	50	24	18	8	M1
x	z	0.5	1.5		2	−1.5		52	23	18	7	M2
x	y, z	0.5	1.5		1.95	−1.5	1.5	51	23	18	8	M3
x, y	y	1.75	1.5	−1.5	−0.5		1.5	50	26	18	6	M4
x, y	z, y	1.75	1.5	−1.5	1	−1.5	1.5	51	25	18	6	M5

* Note: $(x,\cdot)$: subjects with x observed and z missing; $(\cdot,z)$: subjects with x missing and z observed; $(\cdot,\cdot)$: subjects with both x and z missing.

coefficients are used throughout to make bias comparable across all simulation settings;

3. Simulate a binary indicator of response r_z of z from each of the following models.

$$\text{logit}\left[p\left(r_z = 1 \mid y\right)\right] = \alpha_{z0} + \alpha_{zy}y;$$
$$\text{logit}\left[p\left(r_z = 1 \mid z\right)\right] = \alpha_{z0} + \alpha_{zz}z;$$
$$\text{logit}\left[p\left(r_z = 1 \mid z, y\right)\right] = \alpha_{z0} + \alpha_{zz}z + \alpha_{zy}y.$$

4. Simulate a binary indicator of response r_x of x from each of the following models.

$$\text{logit}\left[p\left(r_x = 1 \mid x\right)\right] = \alpha_{x0} + \alpha_{xx}x;$$
$$\text{logit}\left[p\left(r_x = 1 \mid x, y\right)\right] = \alpha_{x0} + \alpha_{xx}x + \alpha_{xy}y.$$

   Five combinations of missing data models for both z and x are considered (M1–M5, table 3.5). Corresponding values of the α_z and α_x parameters are chosen for relatively strong associations in the selection models (odds ratios of 4.48 and 0.22), and are presented in table 3.5. Values of α_{z0} and α_{z0} are altered to achieve the same patterns of missing values in x and z (table 3.5);

5. For $i = 1, \ldots, 100\,000$, set z_i and x_i to missing if $r_{z_i} = 0$ and $r_{x_i} = 0$, respectively;

6. Impute missing values in x and z $M = 10$ times with $T = 10$ iterations using the standard implementation of MICE with all unweighted conditional models for x and z, and marginal and conditional weighted conditional models for x and/or z with p_z^{pop} and p_x^{pop} as reference proportions when the corresponding missingness mechanism is MNAR (i.e. M1–M5 for x and M2, M3, M5 for z).

   The imputation model for x is a logistic regression of x conditional on y, z, and r_z among the observed x. Similarly, the imputation model for z is a logistic regression of z conditional on y, x, and r_x among the observed z. Leacy [80] explored how the standard MICE procedure can be extended for imputation under general MNAR mechanisms by including the δ offsets [23] in the univariate conditional models. It was suggested that if one variable is imputed as MNAR using its response indicator, then its response indicator should also be included in the other imputation models [80]. Therefore, in multivariate simulation studies presented in this thesis, the response indicator of x is included in the imputation model for z, and vice versa;

7. For each MI method, fit the analysis model (3.11) to each completed dataset and combine the results using Rubin's rules [20, 21].

The same full dataset is generated for each missingness scenario, to which data in x and z are set to missing according to the various missingness models considered, and the same incomplete dataset is used to compare the three MICE methods under the same missingness scenario. The parameters of interest are β_{y0}, β_{yx} and β_{yz}. Analyses of the full data and complete records are also provided for comparison.

The algorithm for performing marginal weighted MICE of x and z is as follows.

1. If the population distribution of x/z is known, calculate the marginal weights for x/z from its distribution in the population and among subjects with observed x/z;
2. Fill in missing values in x and z randomly with observed values of x and z, respectively;
3. Begin iteration; for the imputation of x:
   a. Discard the filled-in/imputed values in x;
   b. Fit a (weighted) logistic regression imputation model for x conditional on z (observed

and filled-in/imputed), y (complete), and r_z (complete) to subjects with observed x to obtain maximum likelihood estimates of imputation model's parameters and standard errors;

c. Draw new parameters from the large-sample normal approximation of the posterior distribution, assuming non-informative priors;

d. Generate imputed values from the logistic regression model with new parameters and replace missing values in x with the imputed values;

4. For the imputation of z, follow the same imputation procedure for x. The imputation model for z is a logistic regression of z conditional on x (observed and imputed from the previous step), y (complete), and r_x (complete);
5. Repeat for $T = 10$ iterations to obtain one set of imputed values for x and z.

The algorithm for performing conditional weighted MICE of x and z is as follows.

1. Fill in missing values in x and z randomly with observed values of x and z, respectively;
2. Begin iteration; for the imputation of x:
   a. Discard the filled-in/imputed values in x;
   b. If the population distribution of x is known: fit a logistic regression model for x conditional on z (observed and filled-in/imputed), y (complete), and r_z (complete) to subjects with observed x to get the maximum likelihood estimates of imputation model's parameters and standard errors. These parameter estimates are used to obtain predicted proportions of x in the completed data, which are then used with the population proportions of x to calculate the conditional weights for x.
   c. Fit a (weighted) logistic regression imputation model of x conditional on z (observed and filled-in/imputed), y (complete), and r_z (complete) to subjects with observed x to obtain maximum likelihood estimates of imputation model's parameters and standard errors;
   d. Draw new parameters from the large-sample normal approximation of the posterior distribution, assuming non-informative priors;
   e. Generate imputed values from the logistic regression model with new parameters and replace missing values in x with the imputed values;
3. For the imputation of z, follow the same imputation procedure for x. The imputation model for z is a logistic regression of z conditional on x (observed and imputed from the previous step), y (complete), and r_x (complete);
4. Repeat for $T = 10$ iterations to obtain one set of imputed values for x and z.

All simulations are performed in Stata 14 [44] using `mi impute chained` [75] for standard MICE and my commands for marginal and conditional weighted MICE.

3.6.3 Results

The full-data distributions of y, x, and z are given in table 3.6. Figure 3.10 shows the point estimates of parameters β_{y0}, β_{yx} and β_{yz} of the analysis model after missing values in x and z are imputed using MICE with unweighted conditional models for both x and z, or marginal/conditional weighted conditional models for x and/or z when the corresponding missingness mechanism for x and/or z is MNAR. Overall, the full data analysis yields small bias in the point estimates, with the smallest standard errors. Estimates in CRA have the largest standard errors due to a

Table 3.6. Single multivariate simulation study: distribution of y, x, and z in the full data.

		$x = 0$	$x = 1$	$\sum_x$
$y = 0$	$z = 0$	6 941 (49.81)	6 323 (32.05)	13 264 (39.40)
	$z = 1$	6 993 (50.19)	13 406 (67.95)	20 399 (60.60)
	$\sum_z$	13 934	19 729	**33 663**
$y = 1$	$z = 0$	11 700 (73.07)	28 756 (57.14)	40 456 (60.99)
	$z = 1$	4 312 (26.93)	21 569 (42.86)	25 881 (39.01)
	$\sum_z$	16 012	50 325	**66 337**

* Note: n=100 000; $p\,(y = 1) = 0.66$; cell values are frequency (%).

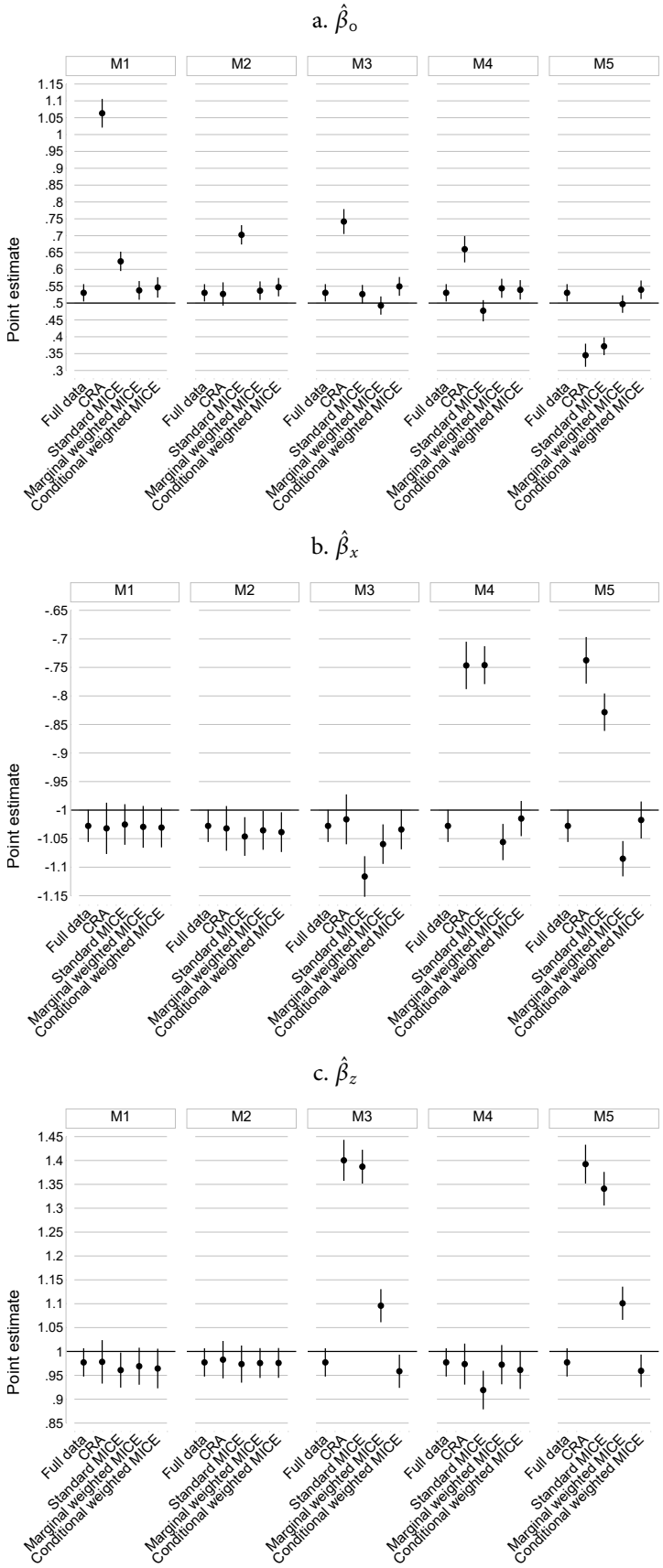
reduction in the sample size by about 50%.

Under M1, when x is MNAR dependent on x and z is MAR conditional on y, CRA yields large bias in the estimate of β_{y0}, but appears to produce estimates that are close to the true values for β_{yx} and β_{yz}. Generally, when the probability of being a complete record depends jointly on x and y, CRA is biased for all parameter estimates including $\hat{\beta}_{yx}$. However, when the missingness mechanism is such that the probability of being a complete record can be written as a product of some suitable function of x and some suitable function of y, CRA is asymptotically unbiased for β_{yx} [79]. Since missing values were generated such that missingness in x is dependent on x and missingness in z is dependent on y, the resulting $\hat{\beta}_{yx}$ is unbiased in CRA. Due to the symmetry of the odds ratio, CRA is also unbiased for $\hat{\beta}_{yz}$. Standard MICE produces smaller bias in the β_{y0} estimate compared to CRA, and the method performs relatively well in terms of bias in the β_{yx} and β_{yz} estimates. MICE with a marginal weighted conditional model for x yields small bias in point estimates, with 95% CIs for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$ covering the true values. Note that 95% CIs in the full data are just about to cover the true values of β_{y0} and β_{yx}. Results in conditional weighted MICE are similar to marginal weighted MICE, with slightly higher bias in $\hat{\beta}_{y0}$.

Under M2, when x is MNAR dependent on x and z is MNAR dependent on z, the probability of being a complete record does not depend on the outcome, conditional on x and z. Therefore, CRA produces unbiased estimates as expected. Results in marginal and conditional weighted MICE are similar to that under M1, with minimal bias in the estimate of β_{y0}. There is also an improvement in the efficiency of MICE compared to CRA, with smaller standard errors and narrower 95% CIs. The efficiency gain is relatively similar for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$, since the percentage of missing values is comparable in x and z. Point estimate of β_{yz} is close to the true value in standard MICE, but the method is biased for $\hat{\beta}_{y0}$ and $\hat{\beta}_{yx}$.

Under M3–M5, when missingness is dependent on both the incomplete variables and the outcome, large bias in point estimates can be seen in CRA and standard MI. Conditional weighted MICE generally produces the smallest bias compared to the other methods across parameter estimates and missingness scenarios considered, and bias tends to be present in $\hat{\beta}_{y0}$ and $\hat{\beta}_{yz}$ in this method. Marginal weighted MICE yields noticeable bias in $\hat{\beta}_{yx}$ under all of these three missingness mechanisms, and in $\hat{\beta}_{yz}$ when z is MNAR (M3 and M5).

Figure 3.10. Single multivariate simulation study: point estimates under different missingness mechanisms for x and z.



* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M3: missingness in x depends on x and in z depends on (z, y); M4: missingness in x depends on (x, y) and in z depends on y; M5: missingness in x depends on (x, y) and in z depends on (z, y); $\beta_{y0} = 0.5$, $\beta_{yx} = -1$, $\beta_{yz} = 1$; horizontal black lines: true parameter values; error bars: 95% confidence intervals.

Table 3.7 presents variance information (including the within- and between-imputation variances ($\widehat{W}$ and $\widehat{B}$, respectively), relative increase in variance (RVI), fraction of missing information (FMI), and relative efficiency (RE), section 2.4.1) about the β_y parameter estimates in the various MICE methods and missingness mechanisms considered for x and z. Within-imputation variances are generally higher than between-imputation variances across methods and missingness mechanisms, and imputation variances are comparable between standard MICE and marginal and conditional weighted MICE. Relative efficiency is above 95% for all parameter estimates and methods.

Table 3.7. Single multivariate simulation study: variance information about the β_y parameters under different missingness mechanisms for x and z.

			$\widehat{W}$	$\widehat{B}$	RVI	FMI	RE
M1	Standard MICE	$\hat{\beta}_{y0}$	0.00019	0.00002	0.120	0.110	0.989
		$\hat{\beta}_{yx}$	0.00022	0.00009	0.479	0.339	0.967
		$\hat{\beta}_{yz}$	0.00023	0.00010	0.496	0.347	0.966
	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.151	0.134	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00012	0.626	0.404	0.961
		$\hat{\beta}_{yz}$	0.00023	0.00013	0.648	0.413	0.960
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00006	0.349	0.270	0.974
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.478	0.339	0.967
		$\hat{\beta}_{yz}$	0.00023	0.00018	0.856	0.485	0.954
M2	Standard MICE	$\hat{\beta}_{y0}$	0.00017	0.00004	0.260	0.214	0.979
		$\hat{\beta}_{yx}$	0.00022	0.00007	0.345	0.267	0.974
		$\hat{\beta}_{yz}$	0.00021	0.00015	0.770	0.457	0.956
	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.143	0.128	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00008	0.423	0.311	0.970
		$\hat{\beta}_{yz}$	0.00023	0.00002	0.111	0.102	0.990
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.129	0.117	0.988
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.481	0.340	0.967
		$\hat{\beta}_{yz}$	0.00023	0.00002	0.115	0.105	0.990
M3	Standard MICE	$\hat{\beta}_{0}$	0.00017	0.00002	0.156	0.139	0.986
		$\hat{\beta}_{yx}$	0.00023	0.00008	0.391	0.293	0.971
		$\hat{\beta}_{yz}$	0.00022	0.00009	0.436	0.318	0.969
	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.151	0.135	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.446	0.321	0.969
		$\hat{\beta}_{yz}$	0.00023	0.00007	0.360	0.276	0.973
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00018	0.00002	0.148	0.132	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.472	0.336	0.968
		$\hat{\beta}_{yz}$	0.00023	0.00008	0.353	0.272	0.974
M4	Standard MICE	$\hat{\beta}_{0}$	0.00018	0.00007	0.411	0.304	0.970
		$\hat{\beta}_{yx}$	0.00021	0.00007	0.374	0.284	0.972
		$\hat{\beta}_{yz}$	0.00022	0.00017	0.824	0.475	0.955
	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00003	0.194	0.167	0.984
		$\hat{\beta}_{yx}$	0.00021	0.00005	0.257	0.211	0.980
		$\hat{\beta}_{yz}$	0.00023	0.00017	0.822	0.474	0.955
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00004	0.237	0.198	0.981
		$\hat{\beta}_{yx}$	0.00021	0.00004	0.194	0.167	0.984
		$\hat{\beta}_{yz}$	0.00023	0.00015	0.730	0.443	0.958
M5	Standard MICE	$\hat{\beta}_{0}$	0.00016	0.00002	0.147	0.131	0.987
		$\hat{\beta}_{yx}$	0.00022	0.00005	0.269	0.220	0.979
		$\hat{\beta}_{yz}$	0.00021	0.00009	0.463	0.331	0.968
	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00016	0.00001	0.070	0.066	0.993
		$\hat{\beta}_{yx}$	0.00021	0.00003	0.174	0.152	0.985
		$\hat{\beta}_{yz}$	0.00023	0.00008	0.363	0.278	0.973
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.119	0.108	0.989
		$\hat{\beta}_{yx}$	0.00021	0.00006	0.306	0.243	0.976
		$\hat{\beta}_{yz}$	0.00023	0.00006	0.301	0.241	0.977

* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M3: missingness in x depends on x and in z depends on (z, y); M4: missingness in x depends on (x, y) and in z depends on y; M5: missingness in x depends on (x, y) and in z depends on (z, y).

3.6.4 *Repeated simulations for assessing performance measures*

This section presents results of a repeated multivariate simulation study conducted to examine other performance measures, when the univariate weighted MI methods are incorporated into MICE for handling multivariate missing data. Repeated simulations are performed for missingness models M1 and M2. Under these missingness mechanisms, marginal weighted MICE produces the least biased parameter estimates out of the three MI methods under comparison, as seen in the single multivariate simulation study (section 3.6.3). M5 is also explored further over repeated simulations since this missingness mechanism is the most complex one, under which missingness in each covariate depends on its values and the outcome.

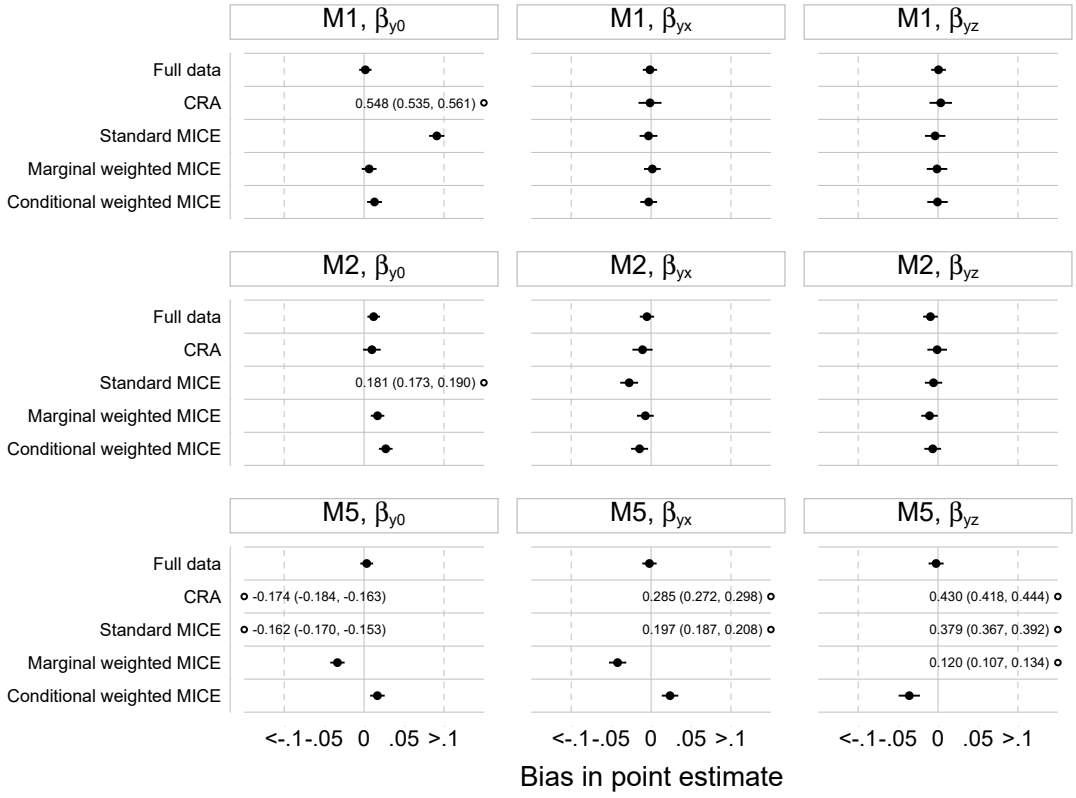
The single simulation set-up outlined in section 3.6.2 for M1, M2, and M5 is performed using $S = 1000$ repetitions and $n = 1000$ observations. A smaller sample size is chosen for repeated simulations to reduce processing time, since it takes relatively longer for the conditional weights to be updated after every iteration in each imputation of the chained equations. In all MICE methods, missing values in x and z are imputed using $M = 10$ imputations and $T = 10$ iterations.

Results are presented graphically in figures 3.11–3.13. Bias over repeated simulations across MICE methods for dealing with missing data and missingness mechanisms considered for x and z is consistent with results in the corresponding large-sample single simulation (figure 3.11). Under M1, marginal and conditional weighted MICE are both unbiased for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$, while there is negligible bias in the estimate of the intercept in conditional weighted MICE. Standard MICE and CRA also yield unbiased estimates of β_{yx} and β_{yz} under this missingness mechanism, but bias is noticeable in the intercept in both methods. Under M2, standard MICE is again unbiased for $\hat{\beta}_{yz}$ but produces large bias in $\hat{\beta}_{y0}$ and small bias in $\hat{\beta}_{yx}$. Minimal bias is also present in $\hat{\beta}_{y0}$ in the full data, as well as in both weighted MICE methods. CRA and marginal weighted MICE are unbiased for the estimates of the two log odds ratios, while there is very small bias in the estimate of β_{yx} in conditional weighted MICE under this missingness mechanism. Under M5, none of the methods are unbiased over repeated simulations. However, conditional weighted MICE appears to be the least biased method, while there is noticeable bias in CRA, standard MICE, and marginal weighted MICE.

Empirical standard errors are generally similar to the average model standard errors across parameters and methods under M1 and M2, with the smallest standard errors seen in the full data, and the largest standard errors in CRA due to a decrease in sample sizes (figure 3.12). The gain in efficiency in the MICE methods is represented by a reduction in the standard errors compared to CRA, particularly in scenarios where both approaches yield unbiased point estimates. Under M5, the empirical standard errors in conditional and marginal weighted MICE no longer match the average model standard errors, which might be due to the presence of bias in point estimates.

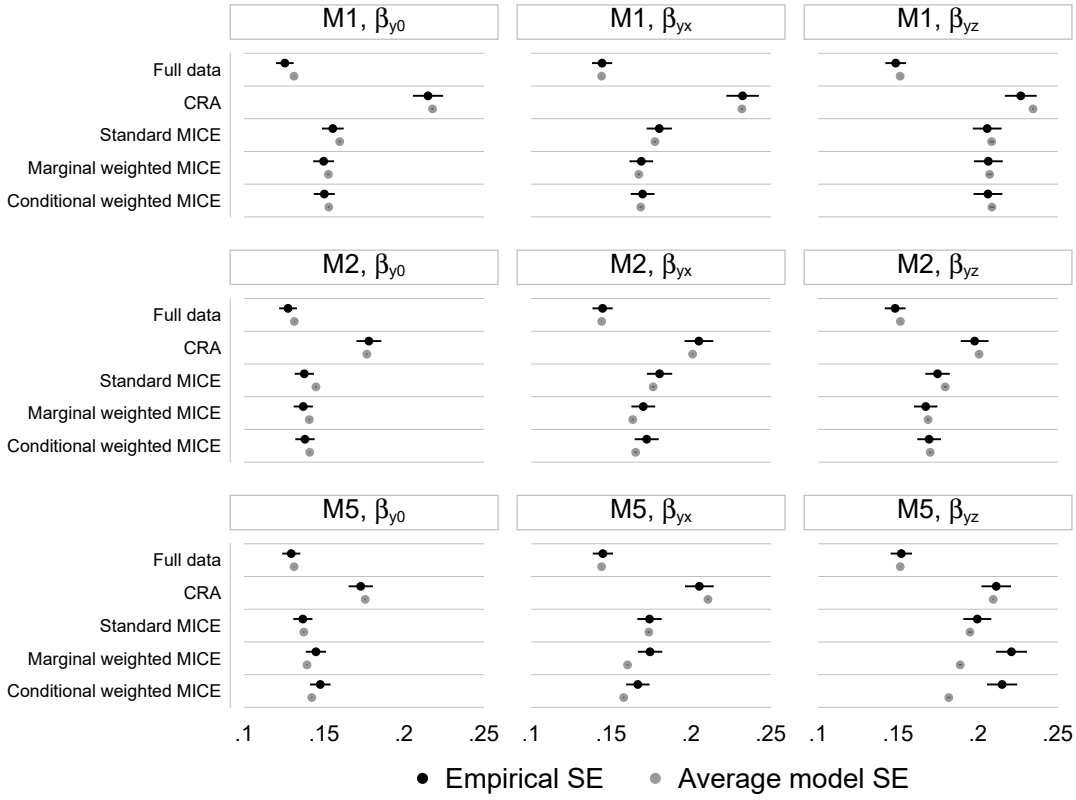
When there is no or very small bias in point estimates, coverage of 95% CIs attains the nominal level (figure 3.13). Coverage is low for all parameters in CRA and standard MICE under M5 when there is substantial bias in point estimates. Although marginal weighted MICE yields noticeable bias in $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$ under M5, coverage remains relatively high at around the 90% level. Conditional weighted MICE achieves the correct coverage under M1 and M2, while there is a reduction in coverage under M5.

Figure 3.11. Repeated multivariate simulation study: bias in point estimates under different missingness mechanisms for x and z.



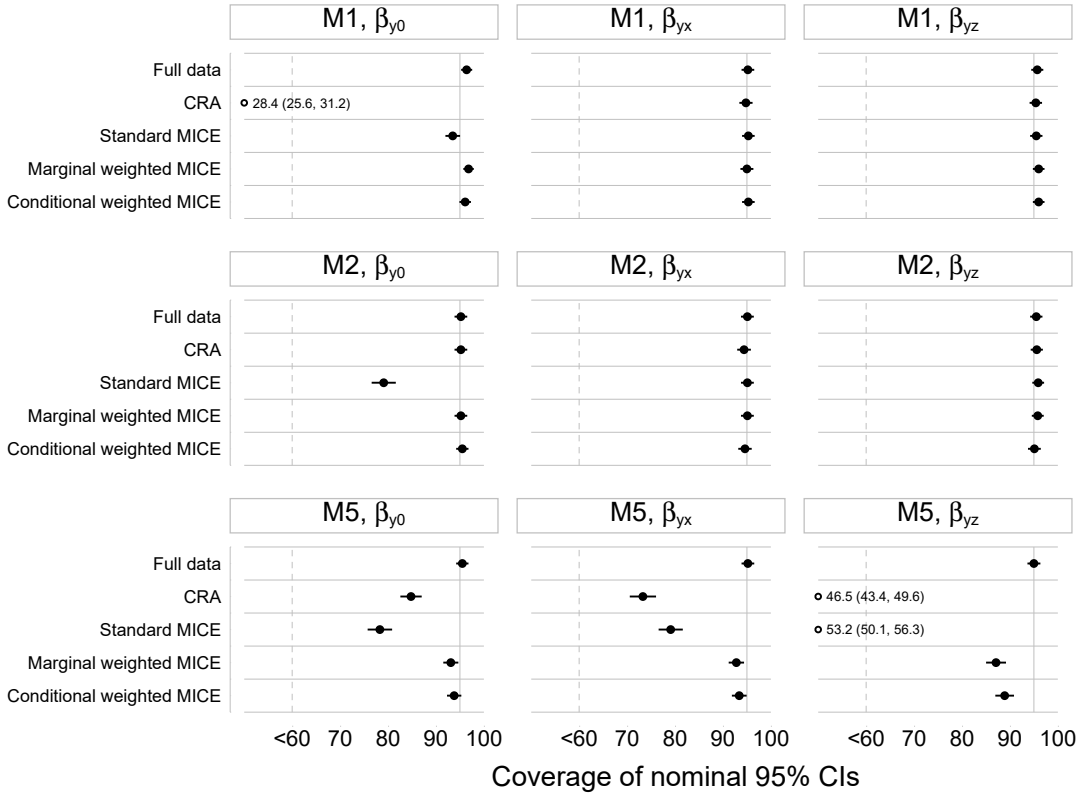
* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); $\beta_0 = 0.5, \beta_x = -1, \beta_z = 1$; error bars: $\pm 1.96\times$ Monte Carlo standard errors; hollow circles: out-of-range values.

Figure 3.12. Repeated multivariate simulation study: empirical and average model standard errors under different missingness mechanisms for x and z.



* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 3.13. Repeated multivariate simulation study: coverage of nominal 95% confidence intervals under different missingness mechanisms for x and z.



* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

3.6.5 *Multivariate simulation studies: conclusion and remarks*

Single and repeated multivariate simulation studies with a complete binary outcome variable and two incomplete binary covariates are conducted to explore the adaptation of univariate marginal and conditional weighted MI in the MICE algorithm for imputing missing values in more than one incomplete covariate.

Results suggest that CRA can be valid in some scenarios. However, even when the method yields unbiased estimates, it still leads to a loss in efficiency with decreased sample sizes and higher standard errors. Apart from M2, standard MICE results in smaller bias in point estimates compared to CRA, but bias is still substantial and coverage is poor.

Marginal weighted MICE appears to be the preferred method when one covariate is MNAR dependent on its values, and the other covariate is either MAR conditional on the outcome (M1) or MNAR dependent on its values (M2). This agrees with findings of the analytic and univariate simulation studies in sections 3.3 and 3.4. When missingness in each covariate depends on both its values and the outcome (M5), conditional weighted MICE appears to produce the smallest bias with relatively high coverage of 95% CIs. This is also consistent with findings of the univariate simulation study in section 3.4. In a 2 × 2 table, when missingness in the incomplete covariate is dependent on its values and the fully observed outcome, conditional weighted MI yields biased estimates, but bias is smaller than that in marginal weighted MI and standard MI (section 3.3.2).

3.7 SUMMARY

This chapter proposes and evaluates the univariate marginal and conditional weighted MI methods, which utilise external population information about the incomplete variable in order to calibrate inference to the population. Throughout this thesis, the focus is on incomplete binary/categorical variables that are included as covariates in the analysis model of interest, and whose population marginal distributions are available in external data sources. The potential of the weighted MI methods for improving on standard MI under general MNAR mechanisms is explored extensively.

In weighted MI, the incomplete variable's population distribution is incorporated as probability weights, which are calculated to match its post-imputation distribution to the reference level. The rationale for the inclusion of probability weights in weighted MI and the formulae for obtaining these weights from the incomplete variable's population distribution are provided in section 3.2.

Weighted MI is evaluated in a univariate missing data setting of a 2×2 contingency table, with a complete outcome variable y and an incomplete covariate x whose population distribution is available externally. Increasingly complex missingness mechanisms are used for generating missing values in x. Section 3.3 presents an analytic study to compare bias in parameter estimates in marginal and conditional weighted MI to that in standard MI and CRA, when the analysis model is a logistic regression of y on x. The weighted MI methods are further evaluated in terms of their frequentist properties in univariate simulation studies in sections 3.4 and 3.5.

Table 3.8. Analytic and univariate simulation studies: summary of bias in the analysis model's parameter estimates under different missingness mechanisms for the incomplete covariate x.

Missingness in x depends on	Method for missing data in x	Biased estimation of	
		β_0	β_x
Neither y nor x	CRA	No	No
	Standard MI	No	No
	Marginal weighted MI	No	No
	Conditional weighted MI	No	No
y	CRA	Yes	No
	Standard MI	No	No
	Marginal weighted MI	Yes	Yes
	Conditional weighted MI	No	No
x	CRA	No	No
	Standard MI	Yes	No
	Marginal weighted MI	No	No
	Conditional weighted MI	Yes	Yes
x and y	CRA	Yes	Yes
	Standard MI	Yes	Yes
	Marginal weighted MI	Yes	Yes
	Conditional weighted MI	Yes	Yes

* Note: analysis model: $\text{logit}\left[p\left(y = 1 \mid x\right)\right] = \beta_0 + \beta_x x$; y (complete) and x (incomplete) are binary variables, taking values 0 or 1.

Results of the analytic and univariate simulation studies are summarised in table 3.8, indicating which method yields unbiased parameter estimates under each of the missingness mechanisms investigated. When the incomplete covariate is MAR conditional on the outcome, conditional weighted MI produces the same results as standard MI and both methods are unbiased. When the incomplete covariate is MNAR dependent on its values, marginal weighted MI results in unbiased point estimates with correct coverage of 95% CIs. However, when the incomplete covariate is MNAR dependent on its values as well as the outcome, both weighted MI methods are not valid, with smaller bias seen in conditional weighted MI.

In section 3.6, the investigation is extended to a multivariate missing data setting to explore the adaptation of the univariate weighted MI methods in the MICE algorithm for multivariate imputation. Multivariate simulation studies are conducted to explore the inclusion of marginal and conditional weighted MI in MICE for imputing missing values in two incomplete binary covariates in a three-way contingency table, when the population distribution of one or both covariates is available externally. In line with the results seen in the univariate missing data setting, marginal weighted MICE produces unbiased parameter estimates when missingness in each of the incomplete covariates depends on its values. In the most complex missingness mechanism considered when each incomplete covariate is missing dependent on its values and the outcome, none of the methods yield unbiased estimates of the analysis model's parameters. However, it is reassuring to see conditional weighted MICE perform well, with relatively small bias in point estimates and good coverage of 95% CIs.

Inspired by van Buuren et al.'s δ adjustment (offset) MI method [23], the next chapter presents an attempt to approach the problem from a different angle, in order to deal with bias seen in marginal and conditional weighted MI when the incomplete covariate is MNAR dependent on its values and the outcome.

4

Calibrated-δ adjustment multiple imputation of a binary covariate when the outcome variable is binary

4.1 INTRODUCTION

This chapter proposes and evaluates *calibrated-δ adjustment multiple imputation* as an alternative approach to weighting (chapter 3) in multiple imputation (MI) when the population-level marginal distribution of the incomplete variable is available externally. In calibrated-δ adjustment MI, knowledge about the incomplete variable's population distribution is utilised to calculate an adjustment in the imputation model's intercept. This intercept adjustment is included in the imputation model in order to tackle bias seen in marginal and conditional weighted MI when missingness in the covariate depends on both its values and the outcome variable (sections 3.3 and 3.4). The idea of the calibrated-δ adjustment is motivated by van Buuren et al.'s δ adjustment (offset) MI method [23]. However, while values of δ are often chosen arbitrarily in van Buuren

et al.'s approach, the population distribution of the incomplete variable is used to derive the value of δ in calibrated-δ adjustment MI to calibrate inference to the population, as the name of the method suggests.

Similar to chapter 3, this chapter also focuses on missing values in an incomplete binary covariate of an analysis model, where the outcome variable is also binary. Featuring a 2×2 contingency table, section 4.2 provides the rationale for calibrated-δ adjustment MI and explains how this method can remove bias in the analysis model's parameter estimates under two missing not at random mechanisms for the covariate. This section also presents the derivation of the calibrated-δ adjustment in the same setting, based on the reference distribution of the incomplete covariate.

Section 4.3 revisits the univariate simulation studies of a 2×2 contingency table with a fully observed binary outcome variable and a partially observed binary covariate, as presented in sections 3.4 and 3.5. Calibrated-δ adjustment MI is evaluated and compared to marginal and conditional weighted MI, standard MI, and complete record analysis in terms of bias in the analysis model's parameter estimates, efficiency, and coverage of 95% confidence intervals, under increasingly complex missingness mechanisms.

In section 4.4, calibrated-δ adjustment MI is further evaluated in a multivariate missing data setting of a three-way contingency table, as outlined in section 3.6. The setting investigated involves a complete binary outcome and two incomplete binary covariates. Repeated simulations are conducted to examine the frequentist properties of calibrated-δ adjustment MI under missing at random (MAR) and missing not at random (MNAR) mechanisms for the covariates.

4.2 THE CALIBRATED-δ ADJUSTMENT MULTIPLE IMPUTATION METHOD

This section begins with an exploration of the equivalence between weighting and including a δ adjustment in the imputation model for an incomplete covariate in a 2×2 contingency table. This exploration is then followed by the derivation of the calibrated-δ adjustment. The frequentist properties of the method is evaluated in simulation studies in subsequent sections.

4.2.1 *An analytic exploration of the equivalence between weighting and δ adjustment in multiple imputation in a 2×2 contingency table*

Recall the example of a 2×2 contingency table in the analytic study presented in section 3.3, where it is of interest to study the association between a binary covariate x taking values $l \in \{0, 1\}$ and a binary outcome y taking values $k \in \{0, 1\}$, with the following analysis model

$$\text{logit}\left[p\left(y = 1 \mid x\right)\right] = \beta_0 + \beta_x x. \tag{4.1}$$

An imputation model

$$\text{logit}\left[p\left(x = 1 \mid y\right)\right] = \theta_0 + \theta_y y \tag{4.2}$$

is fitted to the n_{++}^{obs} complete records to obtain the θ parameter estimates in standard MI, where

$$\hat{\theta}_0^{\text{s}} = \ln\left(\frac{n_{10}^{\text{obs}}}{n_{00}^{\text{obs}}}\right), \hat{\theta}_y^{\text{s}} = \ln\left(\frac{n_{11}^{\text{obs}} n_{00}^{\text{obs}}}{n_{01}^{\text{obs}} n_{10}^{\text{obs}}}\right). \tag{4.3}$$

In weighted MI, the same imputation model is fitted to the weighted complete records,

$n_{lk}^{\text{obs}} w_l^{\text{m/c}}$, where a marginal/conditional weight $w_l^{\text{m/c}}$ is assigned to individuals with observed $x = l$. Parameter estimates of the weighted imputation model are

$$\hat{\theta}_{\text{o}}^{\text{m/c}} = \ln\left(\frac{n_{1\text{o}}^{\text{obs}} w_1^{\text{m/c}}}{n_{0\text{o}}^{\text{obs}} w_{\text{o}}^{\text{m/c}}}\right), \hat{\theta}_y^{\text{m/c}} = \ln\left(\frac{n_{11}^{\text{obs}} n_{0\text{o}}^{\text{obs}}}{n_{01}^{\text{obs}} n_{1\text{o}}^{\text{obs}}}\right). \tag{4.4}$$

Note that in (4.3) and (4.4), the estimated log odds ratio $\hat{\theta}_y$ of the imputation model is the same in standard MI and weighted MI, but the estimated log odds $\hat{\theta}_{\text{o}}$ in weighted MI can be written as $\hat{\theta}_{\text{o}}^{\text{m/c}} = \hat{\theta}_{\text{o}}^{\text{s}} + \ln\left(\frac{w_1}{w_{\text{o}}}\right)$. This implies that in this simple setting, weighted MI is equivalent to the (pattern-mixture) δ adjustment, also known as *offset*, MI method proposed by van Buuren et al. [23]. The imputation model for the incomplete covariate x is therefore

$$\text{logit}\left[p\left(x = 1 \mid y\right)\right] = \theta_{\text{o}} + \theta_y y + \delta\left(1 - r\right),$$

where $\delta = \ln\left(\frac{w_1}{w_{\text{o}}}\right)$ is now estimated using the population marginal distribution of x instead of being chosen arbitrarily. When the incomplete covariate is binary, weighting the imputation model is effectively equivalent to changing the intercept of the imputation model by a factor. This factor needs to be appropriately determined to match the post-imputation distribution of the incomplete covariate to the population level, as well as to preserve the correct outcome–covariate association in the analysis model.

Two MNAR mechanisms considered in the analytic study in section 3.3 (M3 and M4, table 4.1) are now reevaluated. The aim of this investigation is to confirm whether adjusting the intercept of the imputation model can sufficiently recover the correct post-imputation distribution of the incomplete covariate and its association with the outcome.

1. M3 - when x is MNAR dependent on x

   Under this missingness mechanism, the posited model for the response indicator r of x is given by

   $$\text{logit}\left[p\left(r = 1 \mid x\right)\right] = \alpha_{\text{o}} + \alpha_x x, \tag{4.5}$$

   and the corresponding probabilities of observing x are

   $$p\left(r = 1 \mid x = l\right) = p_{r_l} = \text{expit}\left(\alpha_{\text{o}} + \alpha_x x\right).$$

   For imputation model (4.2), the log odds ratios of x for $y = 1$ compared to $y = 0$ in the observed and missing data are given by

   $$\left[\theta_y \mid r = 1\right] = \theta_y^{\text{obs}} = \ln\left(\frac{n_{0\text{o}} p_{r_{\text{o}}} n_{11} p_{r_1}}{n_{01} p_{r_{\text{o}}} n_{1\text{o}} p_{r_1}}\right) = \ln\left(\frac{n_{0\text{o}} n_{11}}{n_{01} n_{1\text{o}}}\right);$$

Table 4.1. Analytic study: models for missingness in x.

Linear predictor of selection model $\text{logit}\left[p\left[\left(r = 1 \mid x, y\right)\right]\right.$	Selection probability $p\left(r_{lk} = 1\right)$	Label
α_{o}	p_r	M1
$\alpha_{\text{o}} + \alpha_y y$	p_{r_k}	M2
$\alpha_{\text{o}} + \alpha_x x$	p_{r_l}	M3
$\alpha_{\text{o}} + \alpha_x x + \alpha_y y$	$p_{r_{lk}}$	M4

* Note: r: response indicator of x; l and k: index categories of x and y, respectively; l, k take values 0 or 1.

$$\left[\theta_y \mid r = 0\right] = \theta_y^{\text{mis}} = \ln\left[\frac{n_{00}\left(1 - p_{r_0}\right) n_{11}\left(1 - p_{r_1}\right)}{n_{01}\left(1 - p_{r_0}\right) n_{10}\left(1 - p_{r_1}\right)}\right] = \ln\left(\frac{n_{00} n_{11}}{n_{01} n_{10}}\right),$$

respectively. Hence, $\theta_y^{\text{obs}} = \theta_y^{\text{mis}}$, which are also the same as the log odds ratio θ_y in the full data.

The log odds of x for $y = 0$ in the observed and missing data are given by

$$\theta_0^{\text{obs}} = \ln\left(\frac{n_{10} p_{r_1}}{n_{00} p_{r_0}}\right);$$
$$\theta_0^{\text{mis}} = \ln\left[\frac{n_{10}\left(1 - p_{r_1}\right)}{n_{00}\left(1 - p_{r_0}\right)}\right],$$

respectively. This implies that the correct adjustment in the imputation model's intercept should be

$$\begin{aligned}\theta_0^{\text{mis}} - \theta_0^{\text{obs}} &= \ln\left[\frac{\left(1 - p_{r_1}\right) p_{r_0}}{\left(1 - p_{r_0}\right) p_{r_1}}\right] \\ &= \ln\left[\frac{\exp\left(\alpha_0\right)}{\exp\left(\alpha_0 + \alpha_x\right)}\right] \\ &= -\alpha_x,\end{aligned}$$

which is minus the log odds ratio of observing x for $x = 1$ compared to $x = 0$ in (4.5).

The log ratio of the two marginal weights (for $x = 0$ and $x = 1$) can be shown to be the same as the correct intercept adjustment, as

$$\begin{aligned}\ln\left(\frac{w_1^m}{w_0^m}\right) &= \frac{\left(n_{1+} - n_{1+}^{\text{obs}}\right) n_{0+}^{\text{obs}}}{n_{1+}^{\text{obs}}\left(n_{0+} - n_{0+}^{\text{obs}}\right)} \\ &= \ln\left[\frac{\left(n_{1+} - n_{1+} p_{r_1}\right) n_{0+} p_{r_0}}{n_{1+} p_{r_1}\left(n_{0+} - n_{0+} p_{r_0}\right)}\right] \\ &= \ln\left[\frac{\left(1 - p_{r_1}\right) p_{r_0}}{\left(1 - p_{r_0}\right) p_{r_1}}\right] \\ &= -\alpha_x.\end{aligned}$$

This result explains the equivalence of marginal weighted MI and calibrated-δ adjustment MI under this missingness mechanism, and why marginal weighted MI provides unbiased parameter estimates of model (4.1) in this case (sections 3.3.3 and 3.4.3).

2. M4 - when x is MNAR dependent on x and y

Under this missingness mechanism, the posited model for the response indicator r of x is given by

$$\text{logit}\left[p\left(r = 1 \mid x, y\right)\right] = \alpha_0 + \alpha_x x + \alpha_y y, \tag{4.6}$$

and the corresponding probabilities of observing x are

$$p\left(r = 1 \mid x = l, y = k\right) = p_{r_{lk}} = \text{expit}\left(\alpha_0 + \alpha_x x + \alpha_y y\right). \tag{4.7}$$

For imputation model (4.2), the log odds ratios of x for $y = 1$ compared to $y = 0$ in the observed and missing data are given by

$$\theta_y^{\text{obs}} = \ln\left(\frac{n_{00} p_{r_{00}} n_{11} p_{r_{11}}}{n_{01} p_{r_{01}} n_{10} p_{r_{10}}}\right); \tag{4.8}$$

$$\theta_y^{\text{mis}} = \ln\left[\frac{n_{00}\left(1-p_{r_{00}}\right)n_{11}\left(1-p_{r_{11}}\right)}{n_{01}\left(1-p_{r_{01}}\right)n_{10}\left(1-p_{r_{10}}\right)}\right]. \tag{4.9}$$

(4.8) and (4.9) can be shown to be equal, since

$$\begin{aligned}
\theta_y^{\text{mis}} - \theta_y^{\text{obs}} &= \ln\left[\frac{\left(1-p_{r_{00}}\right)\left(1-p_{r_{11}}\right)p_{r_{01}}p_{r_{10}}}{\left(1-p_{r_{01}}\right)\left(1-p_{r_{10}}\right)p_{r_{00}}p_{r_{11}}}\right] \\
&= \ln\left[\frac{\exp\left(\alpha_0+\alpha_x\right)\exp\left(\alpha_0+\alpha_y\right)}{\exp\left(\alpha_0\right)\exp\left(\alpha_0+\alpha_x+\alpha_y\right)}\right] \\
&= \ln\left(1\right) \\
&= 0.
\end{aligned}$$

The log odds of x for $y = 0$ in the observed and missing data are given by

$$\begin{aligned}
\theta_0^{\text{obs}} &= \ln\left(\frac{n_{10}p_{r_{10}}}{n_{00}p_{r_{00}}}\right); \\
\theta_0^{\text{mis}} &= \ln\left[\frac{n_{10}\left(1-p_{r_{10}}\right)}{n_{00}\left(1-p_{r_{00}}\right)}\right],
\end{aligned}$$

which implies that the correct adjustment in the intercept of the imputation model should be

$$\begin{aligned}
\theta_0^{\text{mis}} - \theta_0^{\text{obs}} &= \ln\left[\frac{\left(1-p_{r_{10}}\right)p_{r_{00}}}{\left(1-p_{r_{00}}\right)p_{r_{10}}}\right] \\
&= \ln\left[\frac{\exp\left(\alpha_0\right)}{\exp\left(\alpha_0+\alpha_x\right)}\right] \\
&= -\alpha_x.
\end{aligned}$$

This is again minus the log odds ratio of observing x in (4.6). However, the correct intercept adjustment is no longer the same as either the log ratio of the two marginal weights (for $x = 0$ and $x = 1$) or conditional weights. This finding explains bias seen in both marginal and conditional weighted MI under this missingness mechanism (sections 3.3.3 and 3.4.3).

4.2.2 *Derivation of the calibrated-δ adjustment*

The analytic calculations in section 4.2.1 confirm that in a 2×2 contingency table setting, appropriately adjusting the intercept of the imputation model for the covariate x sufficiently corrects bias introduced by MNAR mechanisms under which missingness in x depends on either its values or both its values and the outcome (M3 and M4). By approaching the problem from this angle, the population distribution of x can be used to calculate the correct adjustment in the intercept of the imputation model. This adjustment is referred to as the *calibrated-δ adjustment* to avoid confusion with van Buuren et al.'s δ adjustment.

The probability of $x = 1$ can be written in terms of the conditional probabilities among subjects with observed and missing x

$$p\left(x=1\right) = p\left(x=1 \mid r=1\right)p\left(r=1\right) + p\left(x=1 \mid r=0\right)p\left(r=0\right). \tag{4.10}$$

In (4.10), $p\left(x=1\right)$ is the population proportion; $p\left(x=1 \mid r=1\right)$, $p\left(r=1\right)$, and $p\left(r=0\right)$ are observed. Thus, $p\left(x=1 \mid r=0\right)$ can be solved for as

$$p\left(x=1 \mid r=0\right) = \frac{p\left(x=1\right) - p\left(x=1 \mid r=1\right)p\left(r=1\right)}{p\left(r=0\right)}. \tag{4.11}$$

Note that $p\left(x=1 \mid r=0\right)$ can be further written as

$$\begin{aligned} p\left(x=1 \mid r=0\right) &= \sum_{k=0}^{1} p\left(x=1 \mid y=k, r=0\right) p\left(y=k \mid r=0\right) \\ &= \sum_{k=0}^{1} \operatorname{expit}\left(\theta_0^{\text{mis}} + \theta_y^{\text{mis}} I\left[y=k\right]\right) \frac{n_{+k}^{\text{mis}}}{n_{++}^{\text{mis}}} \\ &= \frac{1}{n_{++}^{\text{mis}}} \sum_{k=0}^{1} \operatorname{expit}\left(\theta_0^{\text{mis}} + \theta_y^{\text{mis}} I\left[y=k\right]\right) n_{+k}^{\text{mis}}, \end{aligned} \tag{4.12}$$

where $I\left[\,\right]$ denotes the indicator function taking values 1 if the statement inside the brackets is true and 0 otherwise; k takes values 0 or 1.

It is shown earlier that when x is MNAR dependent on either x or both x and y, $\theta_y^{\text{obs}} = \theta_y^{\text{mis}}$; (4.12) is therefore equal to

$$\begin{aligned} p\left(x=1 \mid r=0\right) &= \frac{1}{n_{++}^{\text{mis}}} \sum_{k=0}^{1} \operatorname{expit}\left(\theta_0^{\text{mis}} + \theta_y^{\text{obs}} I\left[y=k\right]\right) n_{+k}^{\text{mis}} \\ &= \frac{1}{n_{++}^{\text{mis}}} \sum_{k=0}^{1} \operatorname{expit}\left\{\left(\theta_0^{\text{obs}} + \delta\right) + \theta_y^{\text{obs}} I\left[y=k\right]\right\} n_{+k}^{\text{mis}}, \end{aligned}$$

where δ is the adjustment factor in the intercept of the imputation model for x. Since n_{+k}^{mis} and n_{++}^{mis} are available in the observed data, the value of the calibrated-δ adjustment can be derived from (4.11) and (4.12) using interval bisection [81, 82] (or any other root-finding method).

The interval bisection algorithm for finding the root of an equation involves using the Intermediate Value Theorem [81] to find an initial interval containing the root. In each successive step of the algorithm, the interval is divided in half to get a smaller interval. Eventually, an interval is reached, whose midpoint will be the numerical solution. Suppose f is a continuous function defined on the interval $\left[a_1, a_2\right]$. To find the root of f, interval bisection proceeds as follows.

1. Find a suitable interval $\left[a_1, a_2\right]$ such that $f\left(a_1\right)$ and $f\left(a_2\right)$ are of opposite sign;
2. Calculate the midpoint a_{mid} of the interval, $a_{\text{mid}} = \left(a_1 + a_2\right)/2$;
3. Compute $f\left(a_{\text{mid}}\right)$;
4. If the result in step 3 is sufficiently close to 0, stop iterating and return a_{mid} as the solution;
5. Check the sign of $f\left(a_{\text{mid}}\right)$;
6. Replace a_1 or a_2 with a_{mid} such that the root is still within the new interval;
7. Repeat from step 2.

This approach should yield unbiased estimates of the β parameters under all four missingness mechanisms considered. For M1 and M2, when x is either MCAR or MAR conditional on y, values of the calibrated-δ adjustment derived should be very close to 0. This is because the standard MI approach is unbiased under the MCAR and MAR mechanisms. For MNAR models M3 and M4, calibrated-δ adjustment MI should remove bias seen in both marginal and conditional weighted MI (sections 3.3.3 and 3.4.3).

Note that the derivation of the calibrated-δ adjustment differs from that of the conditional weights in weighted MI, in which the term $p\left(x=1 \mid r=1\right)$ in (4.11) is replaced with

$$p_1^{\text{pred}} = p\left(x=1 \mid y=0, r=1\right) p\left(y=0\right) + p\left(x=1 \mid y=1, r=1\right) p\left(y=1\right).$$

This implies that conditional weighted MI is not the optimal approach for estimating the proba-

bility $p\,(x = 1 \mid r = 0)$ of x in the missing data.

4.3 UNIVARIATE SIMULATION STUDIES – REVISITED

This section presents a univariate simulation study to evaluate performance measures of the calibrated-δ adjustment MI method for an incomplete binary covariate x, when the fully observed outcome variable y is also binary. As before, the aims of this simulation study are to examine finite-sample properties of calibrated-δ adjustment MI in terms of bias in parameter estimates, efficiency, and coverage of 95% confidence intervals (CI); and to compare the method with marginal and conditional weighted MI, standard MI, and complete record analysis (CRA) under various missingness mechanisms for x.

4.3.1 Method

This simulation study is performed using the method outlined in section 3.4.1, with calibrated-δ adjustment MI as an additional method under evaluation.

To implement calibrated-δ adjustment MI of an incomplete covariate x in Stata, first a logistic regression imputation model of x conditional on y is fitted to the complete records and parameter estimates $\hat{\theta}_0^{\text{obs}}$ and $\hat{\theta}_y^{\text{obs}}$ are saved. Next, the calibrated-δ adjustment is obtained numerically using interval bisection as described in section 4.2.2 and stored in a local macro `delta`. The interval bisection algorithm is straightforward to program in Stata. A response indicator r of x is created. The calibrated-δ adjustment is then built into the imputation process via the `offset` option, which is specific to `mi impute logit` [75], using the following commands.

```
. generate offsetvar = -‘delta’*r
. mi impute logit x y, offset(offsetvar) add(50)
```

By executing these commands, a logistic regression imputation model of x conditional on y is fitted to the complete records to obtain maximum likelihood estimates of the imputation model's parameters and their asymptotic sampling variance. The coefficient of the variable `offsetvar` containing the offset is constrained to 1. New parameters are then simulated from the large-sample normal approximation of their posterior distribution, assuming non-informative priors. Imputed values are then generated by drawing from the logistic regression imputation model, given the new parameters.

An equivalence of calibrated-δ adjustment MI, referred to as *calibrated-δ weighted multiple imputation*, is also examined in this simulation study. In this method, the numerical solution of the calibrated-δ adjustment is obtained in the same way, but the adjustment is incorporated into the imputation process as probability weights instead of an offset. Since $\delta = \ln\left(\frac{w_1}{w_0}\right)$ (section 4.2), a weight $w_0 = 1$ can be assigned to subjects with observed $x = 0$, and a weight $w_1 = \exp(\delta)$ to subjects with observed $x = 1$. The imputation can then be performed using the `pweight` option in `mi impute logit` [75], as follows.

```
. generate w = 1 if x == 0
. replace w = exp(‘delta’) if x == 1
. mi impute logit x y [pweight = w]
```

This approach is expected to produce the same results as including the calibrated-δ adjustment as an offset.

Different sets of simulated datasets are generated for each of the four missingness mechanisms considered (M1–M4, table 4.1). This set of missingness mechanisms represent all possible mechanisms for the incomplete covariate in this setting (excluding the interaction between the covariate and the outcome in the selection model for the covariate). Under each missingness mechanism, the full datasets are different across the $S = 1\,000$ simulation repetitions, but the same full dataset is used to compare the various methods for handling missing values in x in each repetition. All simulations are performed in Stata 14 [44], using `mi impute logit` [75] for standard MI, my command `mi impute wlogit`[73] for marginal and conditional weighted MI, `mi impute logit, offset` [75] for calibrated-δ adjustment MI, and `mi impute logit [pweight]` [75] for calibrated-δ weighted MI.

4.3.2 *Results*

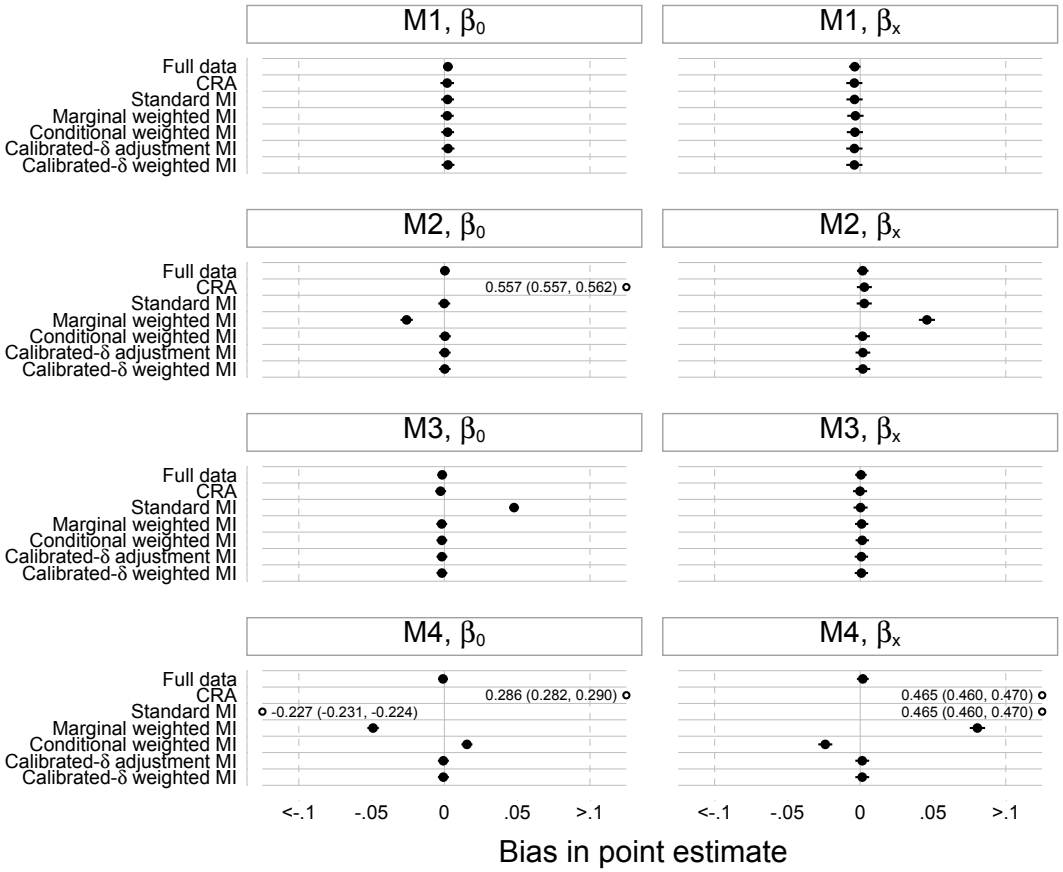
Figures 4.1–4.3 present results of the univariate simulation study. Results in the full data, CRA, marginal and conditional weighted MI are discussed in section 3.4.3 and included here for comparison.

As expected, calibrated-δ adjustment MI and calibrated-δ weighted MI produce identical results in terms of bias, standard errors, and coverage of 95% CIs. This is because in each simulation repetition, the same state of the random-number generator for that repetition is set before performing each of the two MI methods. By resetting the random-number generator states in this way, it is possible to confirm whether these two calibrated-δ MI methods are indeed identical. Both methods are unbiased under all four missingness mechanisms considered, with error bars covering 0 for both $\hat{\beta}_0$ and $\hat{\beta}_x$ (figure 4.1). Bias seen in marginal and conditional weighted MI under M4 is now fully alleviated in calibrated-δ adjustment MI.

Empirical standard errors are similar to the average model standard errors across the four missingness mechanisms in calibrated-δ adjustment MI (figure 4.2). This method produces higher standard errors than that in the full data, as expected. Standard errors of calibrated-δ adjustment MI are comparable to that in standard MI and conditional weighted MI under M2, when the two latter MI methods are unbiased. Standard errors of calibrated-δ adjustment MI are also similar to that in marginal weighted MI under M3, when marginal weighted MI is unbiased.

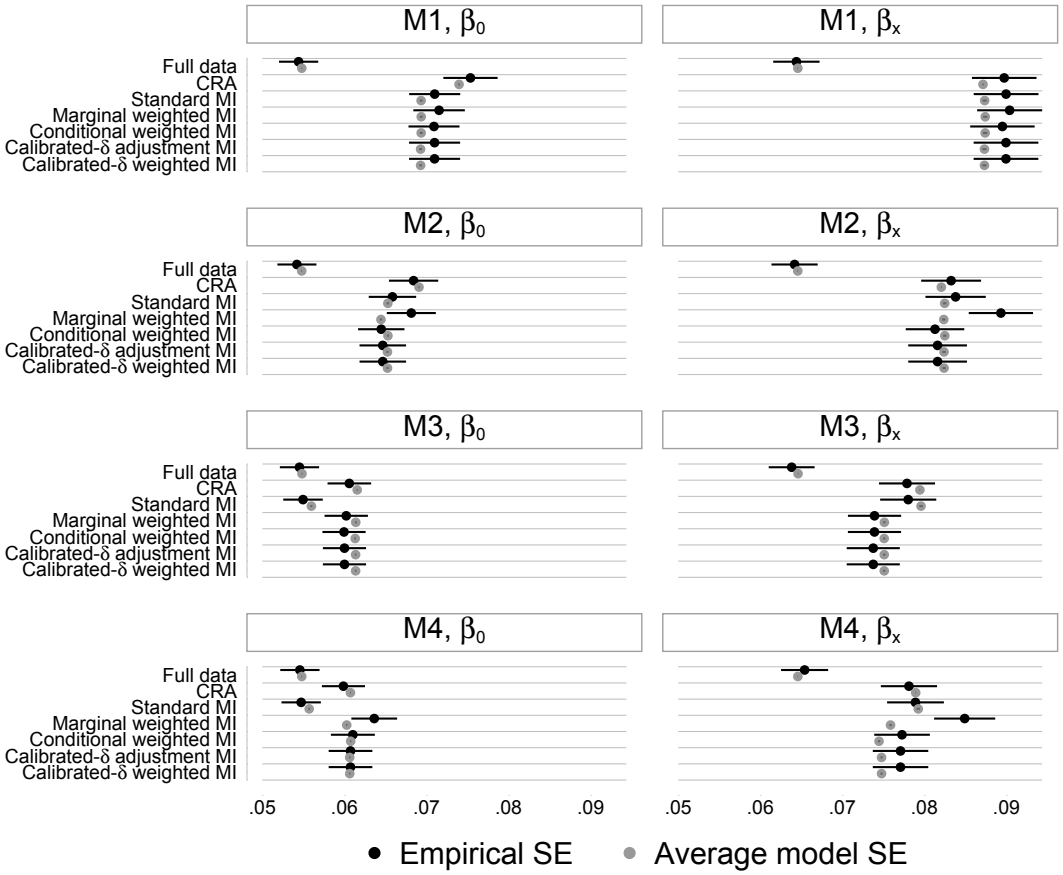
Results for coverage are consistent with that obtained for bias and standard errors. Coverage of 95% CIs in calibrated-δ adjustment MI attains the nominal level for both parameter estimates and under all four missingness mechanisms considered (figure 4.3).

Figure 4.1. Univariate simulation study: bias in point estimates under different missingness mechanisms for x.



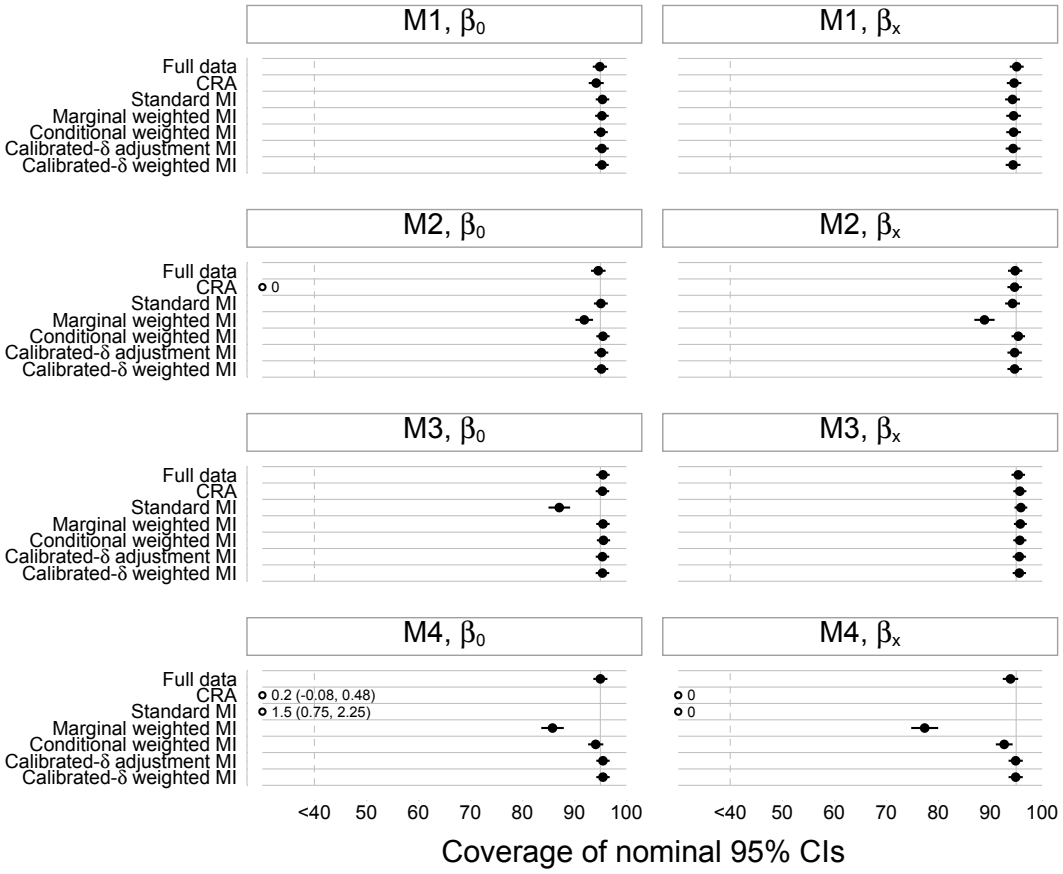
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); $\beta_0 = -0.693$, $\beta_x = 0.405$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure 4.2. Univariate simulation study: empirical and average model standard errors under different missingness mechanisms for x.



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 4.3. Univariate simulation study: coverage of nominal 95% confidence intervals under different missingness mechanisms for x.



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

4.3.3 *Extended univariate simulation study: when there is uncertainty in estimating the population distribution*

The extended univariate simulation study presented in section 3.5 is rerun to examine performance measures of calibrated-δ adjustment MI when the population marginal distribution of the incomplete variable is obtained from an external dataset that is not equivalent to a census. As before, three cases are considered, where the population distribution either comes from a census and is invariant (case 1), or is estimated in external datasets of size $n^{\text{ex}} = 10\,000$ (case 2), and 1 000 (case 3).

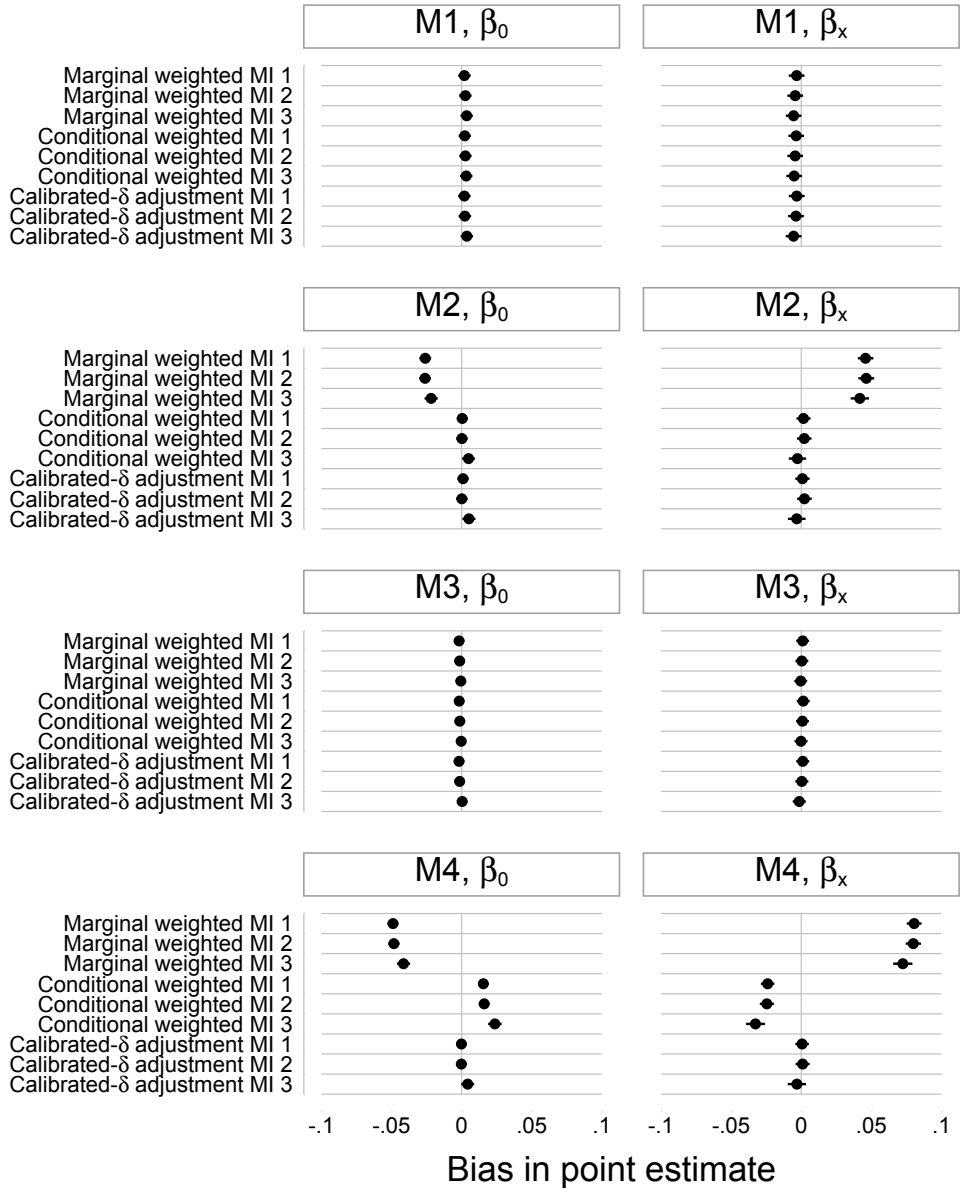
The simulation procedures are the same as that described in section 3.5.1, with a few alterations. For each of the $S = 1\,000$ repetitions, an external dataset of size $n^{\text{ex}} = 10\,000$ or 1 000 is generated in case 2 or 3, respectively, to obtain the estimated population proportion $\hat{p}_x^{\text{pop}}$ of x. Next, a full-data sample of size $n = 5\,000$ is generated for a binary covariate x from a Bernoulli distribution, and a binary outcome y from a logistic regression of y conditional on x. Values of x are then made missing according to selection models M1–M4 (table 4.1). Missing data in x are then imputed by calibrated-δ adjustment MI using $M = 10$ imputations instead of $M = 50$ imputations as before, since the interval bisection algorithm for estimating the calibrated-δ adjustment can take

a relatively long time. In each imputation, a draw of $\dot{p}_x^{\text{pop}}$ is taken from the normal approximation $\text{N}\left(\hat{p}_x^{\text{pop}}, \frac{\hat{p}_x^{\text{pop}}\left(1-\hat{p}_x^{\text{pop}}\right)}{n^{\text{ex}}}\right)$. The calibrated-$\delta$ adjustment is then estimated using $\dot{p}_x^{\text{pop}}$ as the reference proportion. The analysis model, which is a logistic regression of y on x, is then fitted to each completed dataset and the results are combined using Rubin's rules [20, 21]. All simulations are performed in Stata 14 [44].

Figures 4.4–4.6 present the simulation results for calibrated-δ adjustment MI. Results for marginal and conditional weighted MI are also included for reference.

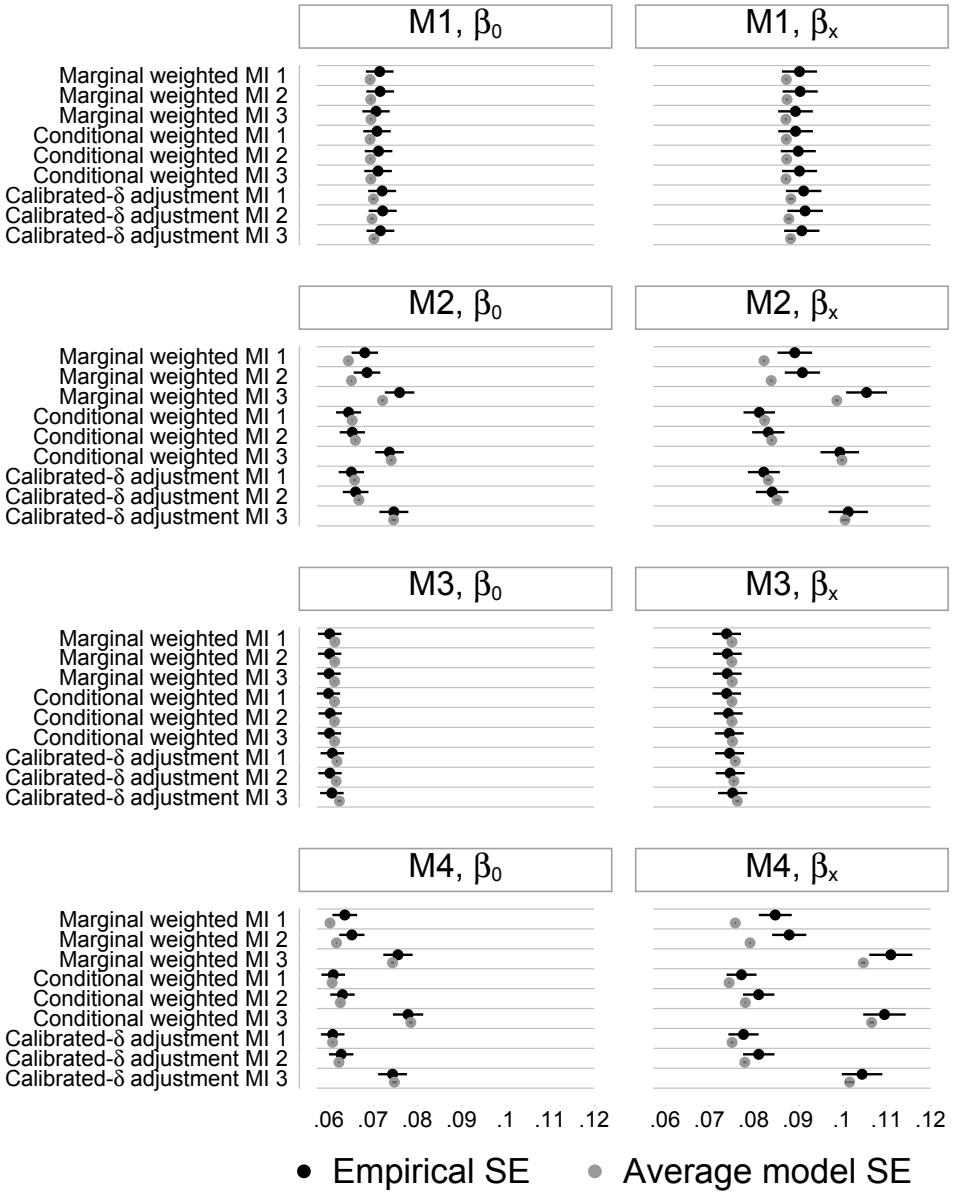
Overall, the results seen in calibrated-δ adjustment MI are similar to that discussed in section 3.5.2. Calibrated-δ adjustment MI appears unbiased in all cases considered. Compared to case 1 where the population distribution of x is invariant, bias in point estimates slightly deviates from 0 with the extra uncertainty in estimating p_x^{pop} in case 3 (figure 4.4). The average model standard errors in calibrated-δ adjustment MI increase markedly in case 3, which is matched by an increase in the corresponding empirical standard errors. These trends are also seen for marginal and conditional weighted MI (figure 4.5). In contrast to marginal weighted MI, coverage in calibrated-δ adjustment MI does not appear to be much affected by the extra uncertainty coming from estimating p_x^{pop} (figure 4.6).

Figure 4.4. Extended univariate simulation study: bias in point estimates under different missingness mechanisms for x; the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).



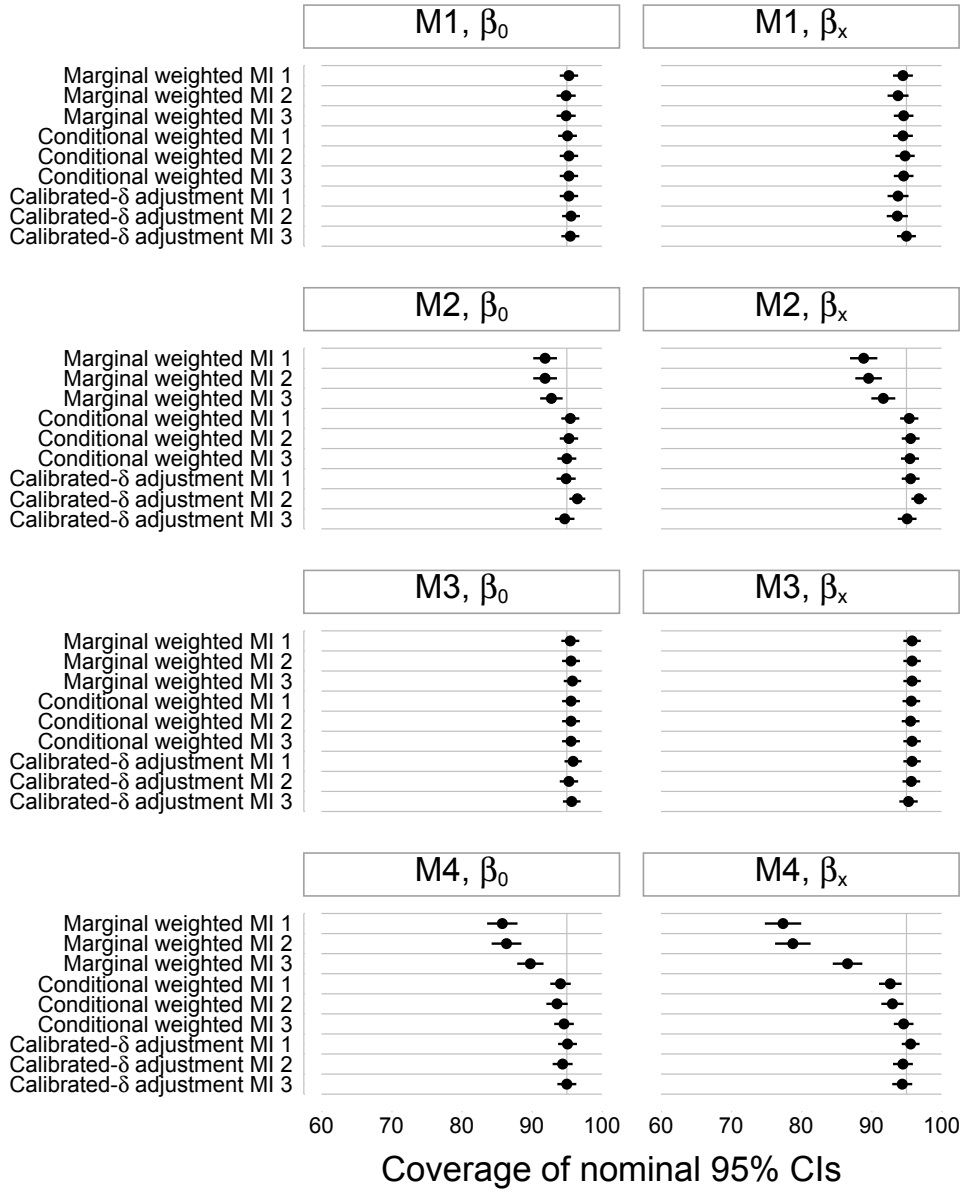
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); $\beta_0 = -0.693$, $\beta_x = 0.405$; error bars: ±1.96× Monte Carlo standard errors.

Figure 4.5. Extended univariate simulation study: empirical and average model standard errors under different missingness mechanisms for x; the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 4.6. Extended univariate simulation study: coverage of nominal 95% confidence intervals under different missingness mechanisms for x; the population distribution of x is assumed to be invariant (case 1) or estimated in external datasets of sizes 10 000 (case 2) and 1 000 (case 3).



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

4.3.4 *Univariate simulation studies: conclusion and remarks*

Results of the univariate simulation studies in sections 4.3.2 and 4.3.3 support the findings in the analytic calculations which explore the equivalence of weighting and including a δ adjustment in MI in a 2×2 contingency table (section 4.2.1). When both the outcome variable y and the covariate x are binary, appropriately adjusting the intercept of the imputation model for x is sufficient to correct bias in point estimates introduced by MNAR mechanisms where missingness in x depends on either the values of x (M3) or both x and y (M4). Having established this, the calibrated-δ adjustment MI method is then equivalent to standard MI when x is MCAR or MAR conditional on y. Calibrated-δ adjustment MI produces the same inferences as marginal weighted MI when x is MNAR dependent on x. Calibrated-δ adjustment MI is unbiased with comparable empirical and average model standard errors and correct coverage of 95% CIs across the four missingness mechanisms considered. In addition, bias previously seen in marginal and conditional weighted MI under M4 is fully removed by the inclusion of the calibrated-δ adjustment.

It is also shown that including the calibrated-δ adjustment as an offset in the logistic regression imputation model yields the same results as incorporating it in the imputation process in the form of probability weights. This has a practical implication on the inclusion of the calibrated-δ adjustment in a multinomial logistic regression imputation model for an incomplete categorical covariate, since the implementation of `mi impute mlogit` in Stata does not support an `offset` option [75]. Under the four missingness mechanisms considered, it is concluded that calibrated-δ adjustment MI is the method of choice for imputing missing values in a binary covariate when the fully observed outcome variable is also binary.

Calibrated-δ adjustment MI is evaluated in analytic and univariate simulation studies of the same set-up as described in sections 3.3–3.5. Using the same reasoning as presented in section 3.5.3, results in this setting can be generalised to the case where the incomplete covariate is a categorical variable with more than two levels. As before, suppose that the covariate contains L levels, and the imputation model is a multinomial logistic regression model with $L - 1$ logistic regression equations contrasting each of the $l = 1, \ldots, L-1$ levels with the base level. The population marginal distribution of the incomplete covariate can be used to derive $L - 1$ appropriate calibrated-δ adjustments in the intercept of each of these regression equations. This helps to justify the use of calibrated-δ adjustment MI (as well as marginal and conditional weighted MI) in two case studies (sections 6.4 and 6.5) examining the issue of missing ethnicity data in UK primary care databases, where ethnicity is analysed as a four-level categorical variable. Further simulations can be undertaken to confirm this generalisability.

4.4 Multivariate simulation studies – revisited

This section explores the integration of the univariate calibrated-δ adjustment MI method in the multivariate imputation by chained equations (MICE) algorithm [23] (section 2.4.3) for imputing missing values in several incomplete covariates. This extension is referred to as *calibrated-δ adjustment MICE*. In the Stata command `mi impute chained` which performs multivariate imputation by chained equations, users can specify a logistic regression conditional model for

an incomplete binary variable with the option `offset` [75]. This is illustrated below for the incomplete variable z_j, $j = 1 \dots q$, at iteration $t + 1$.

$$
\begin{aligned}
& z_1^{(t+1)} \mid z_2^{(t)}, z_3^{(t)}, \dots, z_q^{(t)} && \longleftarrow \text{offset} = \delta_{z_1}; \\
& z_2^{(t+1)} \mid z_1^{(t+1)}, z_3^{(t)}, \dots, z_q^{(t)} && \longleftarrow \text{offset} = 0; \\
& \vdots \\
& z_q^{(t+1)} \mid z_1^{(t+1)}, z_2^{(t+1)}, \dots, z_{q-1}^{(t+1)} && \longleftarrow \text{offset} = 0.
\end{aligned}
$$

However, the offset included in the logistic regression conditional model as implemented in `mi impute chained (logit, offset)` is fixed across all imputations and iterations. Based on the Stata community-contributed command `uvis` [74], I have written a new Stata command to perform multivariate imputation by chained equations. This command allows for an offset specification option in each logistic regression conditional model, such that the offset is calculated in each imputation and updated after every iteration of the algorithm. The purpose of this feature is to account for the fact that missing values are randomly filled in using observed values at the beginning of the iterative process in each imputation. In addition, after each iteration in a given imputation, the incomplete variable is imputed conditional on observed and imputed values of other variables from the previous iteration, which needs to be reflected in the calculation of the calibrated-δ adjustment.

The following sections report a single multivariate simulation study in which the analysis model is a logistic regression of a fully observed binary outcome on two incomplete binary covariates. The aims of this study are to evaluate performance measures of calibrated-δ adjustment MICE; and to compare the method to marginal and conditional weighted MICE and standard MICE under different missingness mechanisms for the covariates.

4.4.1 *Method*

The single multivariate simulation study presented in this section follows the same set-up as outlined in section 3.6.2. To recap, a single large dataset of size $n = 100\,000$ is simulated for a binary outcome y and two binary covariates x and z, to demonstrate large-sample bias in the

Table 4.2. Single multivariate simulation study: variables associated with missingness in x and z, corresponding selection parameters, and percentages of observed data in x and z.

Variables associated with missingness		Selection parameters						% observed data				Label
x	z	α_{x0}	α_{xx}	α_{xy}	α_{z0}	α_{zz}	α_{zy}	(x,z)	$(x,\cdot)$	$(\cdot,z)$	$(\cdot,\cdot)$	
x	y	0.5	1.5		−0.15		1.5	50	24	18	8	M1
x	z	0.5	1.5		2	−1.5		52	23	18	7	M2
x	y, z	0.5	1.5		1.95	−1.5	1.5	51	23	18	8	M3
x, y	y	1.75	1.5	−1.5	−0.5		1.5	50	26	18	6	M4
x, y	z, y	1.75	1.5	−1.5	1	−1.5	1.5	51	25	18	6	M5

* Note: $(x,\cdot)$: subjects with x observed and z missing; $(\cdot,z)$: subjects with x missing and z observed; $(\cdot,\cdot)$: subjects with both x and z missing.

point estimates if present. Values of x and z are made missing under the various missingness mechanisms M1–M5 (table 4.2). The calibrated-δ adjustment is incorporated in the logistic regression conditional models used to impute missing values in x and z in the chained equations, whenever the corresponding missingness mechanism is MNAR, i.e. M1–M5 for x and M2, M3, M5 for z. The algorithm is performed using $M = 10$ imputations with $T = 10$ iterations. The analysis model is then fitted to each completed dataset and the results are combined using Rubin's rules [20, 21].

The algorithm for performing calibrated-δ adjustment MICE for x and z is as follows.

1. Fill in missing values in x and z randomly with observed values of x and z, respectively;
2. Begin iteration; for the imputation of x:
   a. Discard the filled-in/imputed values of x;
   b. Fit a logistic regression model for x conditional on z (observed and filled-in/imputed), y (complete) and r_z (complete) to subjects with observed x to obtain maximum likelihood estimates of the imputation model's parameters and associated standard errors;
   c. Calculate the calibrated-δ adjustment in the intercept of the conditional imputation model for x using interval bisection, as outlined in section 4.2.2. The probability of $x = 1$ among the missing x can be written as

$$p\left(x = 1 \mid r = 0\right) = \frac{1}{n_x^{\text{mis}}} \sum_{i=1}^{n_x^{\text{mis}}} \text{expit}\left[\left(\theta_{x\text{o}}^{\text{obs}} + \delta_x\right) + \theta_{xy}^{\text{obs}} y_i + \theta_{xz}^{\text{obs}} z_i + \theta_{xr_z}^{\text{obs}} r_{z_i}\right],$$

   where i indexes subjects in the dataset, and the estimate of θ_{xr_z} is expected to be close to 0, since conditional on y and z missingness in z does not depend on x;
   d. Fit a logistic regression model for x conditional on z (observed and filled-in/imputed), y (complete) and r_z (complete), with the calibrated-δ adjustment fixed as an offset, to subjects with observed x to obtain maximum likelihood estimates of the imputation model's parameters and associated standard errors;
   e. Draw new parameters (keeping calibrated-δ fixed) from the large-sample normal approximation of their posterior distribution, assuming non-informative priors;
   f. Generate imputed values from the logistic regression model with new parameters and replace missing values in x with the imputed values;
3. For the imputation of z, follow the same imputation procedure for x. The imputation model for z is a logistic regression of z conditional on x (observed and imputed from the previous step), y (complete), and r_x (complete);
4. Repeat for $T = 10$ iterations to obtain one set of imputed values for x and z.

4.4.2 *Results*

Figures 4.7 and 4.8 present the results of the multivariate simulation study. Results in the full data, CRA, standard MICE, and marginal and conditional weighted MICE are discussed in section 3.6.3 and included here for comparison.

Overall, calibrated-δ adjustment MICE gives point estimates that are closest to the full data (figure 4.7). Point estimates are generally closest to the true values in the full data, with the smallest standard errors and the narrowest 95% CIs. CRA has the largest standard errors and the longest 95% CIs. Across the five missingness mechanisms considered, point estimates in

calibrated-δ adjustment MICE are generally similar to that in the full data and conditional weighted MICE, with error bars crossing the true values in most cases for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$. The 95% CIs in calibrated-δ adjustment MICE just about cover the true value of β_{y0} under all five missingness mechanisms, which is similar to the results in conditional weighted MICE and the full data.

Figure 4.8 presents a comparison of the marginal and conditional weights and calibrated-δ adjustment under the five missingness mechanisms used for generating missingness in x and z. The log ratio of the two marginal or conditional weights $\ln\left(\frac{w_1^{\mathrm{m/c}}}{w_0^{\mathrm{m/c}}}\right)$ is plotted against the calibrated-δ adjustment in each of the $M = 10$ imputations for x (under M1–M5) and z (under M2, M3, M5). The conditional weights and calibrated-δ adjustment are taken from the last iteration in each imputation. In addition, the following logistic regression conditional models for x and z

$$\begin{aligned}\operatorname{logit}\left[p\left(x = 1 \mid y, z, r_x\right)\right] &= \theta_{x0} + \theta_{xy}y + \theta_{xz}z + \theta_{xr_x}r_x;\\ \operatorname{logit}\left[p\left(z = 1 \mid y, x, r_z\right)\right] &= \theta_{z0} + \theta_{zy}y + \theta_{zx}x + \theta_{zr_z}r_z,\end{aligned}$$

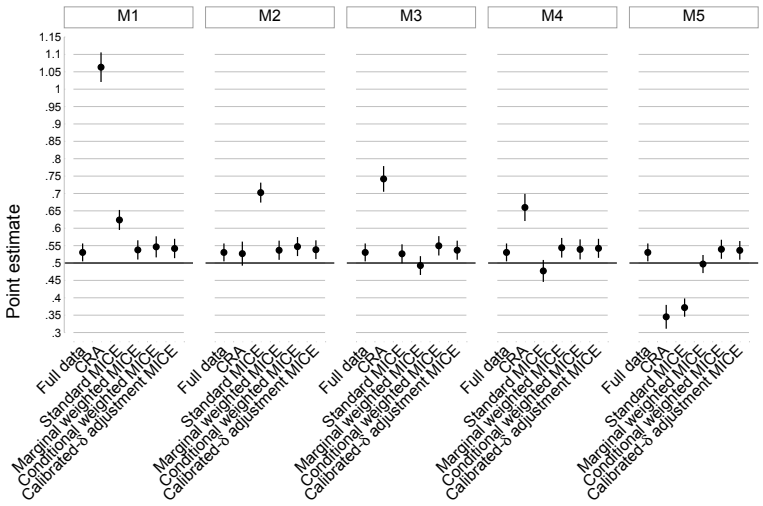
are fitted to the full data (i.e. before any values of x and z are set to missing), and the full-data estimates $\hat{\theta}_{xr_x}^{\mathrm{full}}$ and $\hat{\theta}_{zr_z}^{\mathrm{full}}$ are plotted against the weights and calibrated-δ adjustment for reference.

The values of the calibrated-δ adjustment are close to $\hat{\theta}_{xr_x}$ and $\hat{\theta}_{zr_z}$ for both x and z under all five missingness mechanisms, which agrees with the results for point estimates in calibrated-δ adjustment MICE (figure 4.8). The log ratio of the two marginal weights is similar to $\hat{\theta}_{xr_x}$ and $\hat{\theta}_{zr_z}$ under M1 and M2. This is consistent with the results in figure 4.7 suggesting that point estimates in marginal weighted MICE are comparable to the true values under these two missingness models. Under M3, while the log ratio of the two marginal weights for x is close to $\hat{\theta}_{xr_x}$, the corresponding quantity for z is much smaller in magnitude compared to $\hat{\theta}_{zr_z}$, which explains the noticeable bias in $\hat{\beta}_{yz}$ in marginal weighted MICE. A similar explanation applies to the results in marginal weighted MICE under M4 and M5. Overall, the results in this comparison are reflected in the point estimates for marginal weighted MICE and calibrated-δ adjustment MICE. Therefore, the discrepancy between the log ratio of the conditional weights and $\hat{\theta}_{xr_x}$ and $\hat{\theta}_{zr_z}$ across all missingness mechanisms suggests that some bias is expected to be present in the point estimates in conditional weighted MICE. This is contrary to the results seen in figure 4.7, in which $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$ in conditional weighted MICE are close to the true values. However, bias in conditional weighted MICE is shown over repeated multivariate simulations in section 3.6.4 (figure 3.11).

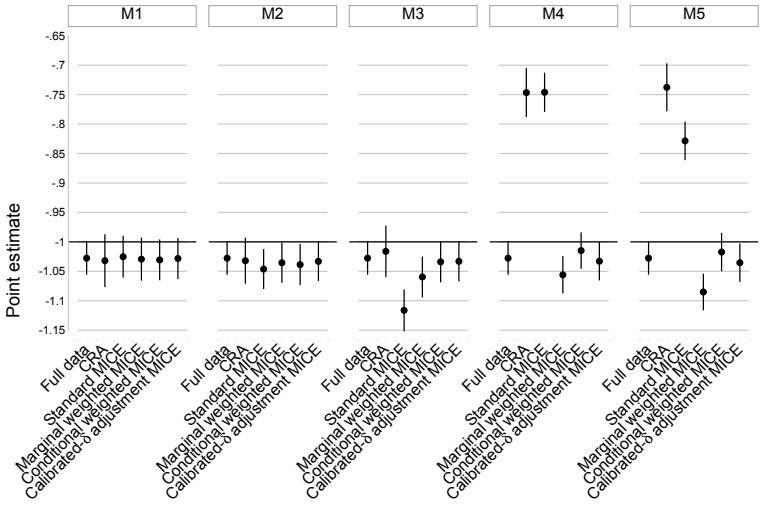
Variance information (including the within- and between-imputation variances ($\widehat{W}$ and $\widehat{B}$, respectively), relative increase in variance (RVI), fraction of missing information (FMI), and relative efficiency (RE), section 2.4.1) about the β parameters in marginal and conditional weighted MICE and calibrated-δ adjustment MICE is presented in table 4.3. The within-imputation variance is stable across the methods and missingness mechanisms under evaluation. Compared to the weighted MICE approaches, calibrated-δ adjustment MICE generally results in similar between-imputation variance for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$. Relative efficiency exceeds 95% for all parameter estimates and methods.

Figure 4.7. Single multivariate simulation study: point estimates under different missingness mechanisms for x and z.

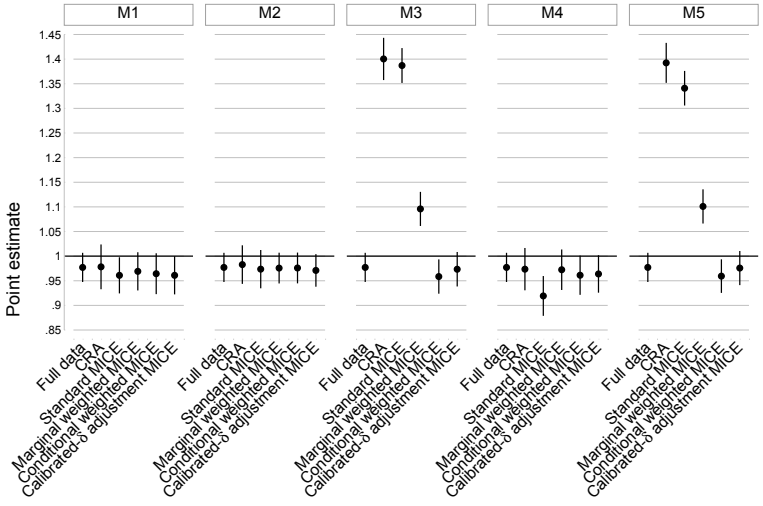
a. $\hat{\beta}_0$



b. $\hat{\beta}_x$

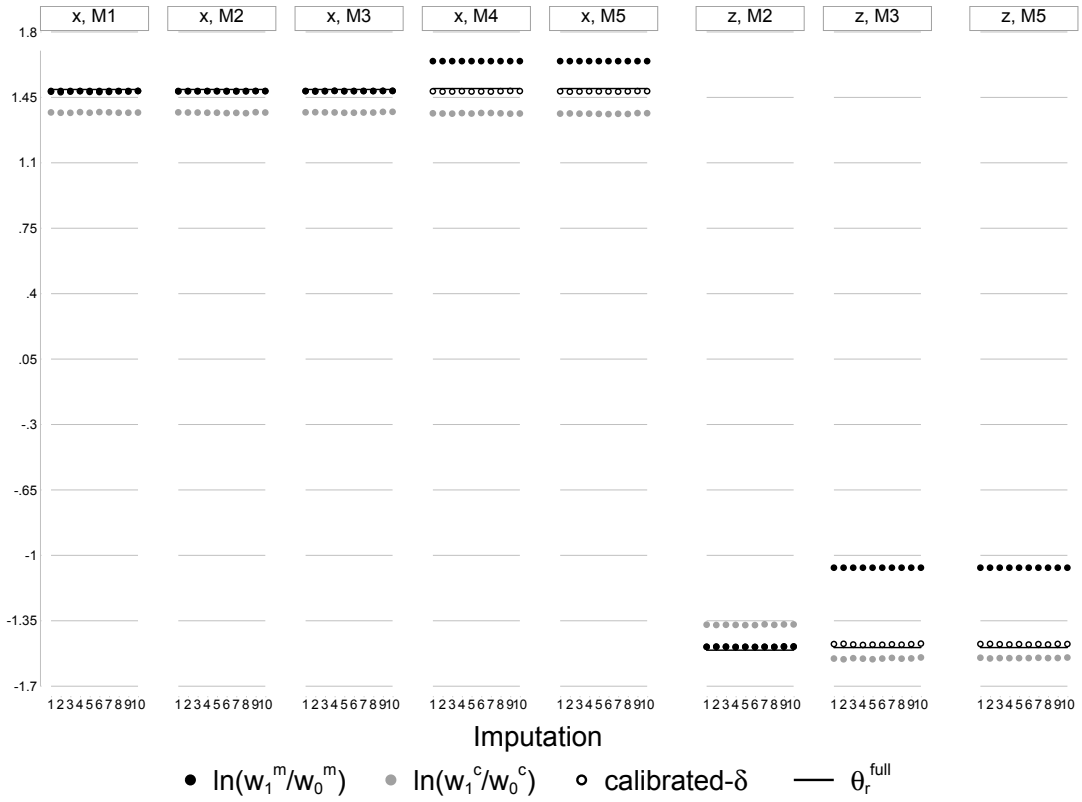


c. $\hat{\beta}_z$



* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M3: missingness in x depends on x and in z depends on (z, y); M4: missingness in x depends on (x, y) and in z depends on y; M5: missingness in x depends on (x, y) and in z depends on (z, y); $\beta_{y0} = 0.5$, $\beta_{yx} = -1$, $\beta_{yz} = 1$; horizontal black lines: true parameter values; error bars: 95% confidence intervals.

Figure 4.8. Single multivariate simulation study: comparison of the marginal and conditional weights, calibrated-δ adjustment, and estimated coefficient of the response indicator across M = 10 imputations.



* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M3: missingness in x depends on x and in z depends on (z, y); M4: missingness in x depends on (x, y) and in z depends on y; M5: missingness in x depends on (x, y) and in z depends on (z, y); the conditional weights and calibrated-δ adjustment are taken from the last iteration of each imputation; for each covariate, a logistic regression model conditional on the outcome, the other covariate, and the covariate's response indicator is fitted to the full data to obtain estimated coefficient of the response indicator $\hat{\theta}_r^{\text{full}}$.

Table 4.3. Single multivariate simulation study: variance information about the β_y parameters under different missingness mechanisms for x and z.

			$\widehat{W}$	$\widehat{B}$	RVI	FMI	RE
M1	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.151	0.134	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00012	0.626	0.404	0.961
		$\hat{\beta}_{yz}$	0.00023	0.00013	0.648	0.413	0.960
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00006	0.349	0.270	0.974
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.478	0.339	0.967
		$\hat{\beta}_{yz}$	0.00023	0.00018	0.856	0.485	0.954
	Calibrated-δ adjustment MICE	$\hat{\beta}_{y0}$	0.00017	0.00003	0.161	0.143	0.986
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.485	0.342	0.967
		$\hat{\beta}_{yz}$	0.00023	0.00013	0.644	0.412	0.960
M2	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.143	0.128	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00008	0.423	0.311	0.970
		$\hat{\beta}_{yz}$	0.00023	0.00002	0.111	0.102	0.990
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.129	0.117	0.988
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.481	0.340	0.967
		$\hat{\beta}_{yz}$	0.00023	0.00002	0.115	0.105	0.990
	Calibrated-δ adjustment MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.098	0.091	0.991
		$\hat{\beta}_{yx}$	0.00021	0.00007	0.376	0.285	0.972
		$\hat{\beta}_{yz}$	0.00023	0.00005	0.235	0.197	0.981
M3	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.151	0.135	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.446	0.321	0.969
		$\hat{\beta}_{yz}$	0.00023	0.00007	0.360	0.276	0.973
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00018	0.00002	0.148	0.132	0.987
		$\hat{\beta}_{yx}$	0.00021	0.00009	0.472	0.336	0.968
		$\hat{\beta}_{yz}$	0.00023	0.00008	0.353	0.272	0.974
	Calibrated-δ adjustment MICE	$\hat{\beta}_{0}$	0.00017	0.00002	0.129	0.117	0.988
		$\hat{\beta}_{yx}$	0.00021	0.00008	0.419	0.308	0.970
		$\hat{\beta}_{yz}$	0.00023	0.00008	0.372	0.283	0.973
M4	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00003	0.194	0.167	0.984
		$\hat{\beta}_{yx}$	0.00021	0.00005	0.257	0.211	0.980
		$\hat{\beta}_{yz}$	0.00023	0.00017	0.822	0.474	0.955
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00004	0.237	0.198	0.981
		$\hat{\beta}_{yx}$	0.00021	0.00004	0.194	0.167	0.984
		$\hat{\beta}_{yz}$	0.00023	0.00015	0.730	0.443	0.958
	Calibrated-δ adjustment MICE	$\hat{\beta}_{0}$	0.00017	0.00002	0.123	0.112	0.989
		$\hat{\beta}_{yx}$	0.00021	0.00006	0.321	0.253	0.975
		$\hat{\beta}_{yz}$	0.00023	0.00012	0.589	0.389	0.963
M5	Marginal weighted MICE	$\hat{\beta}_{y0}$	0.00016	0.00001	0.070	0.066	0.993
		$\hat{\beta}_{yx}$	0.00021	0.00003	0.174	0.152	0.985
		$\hat{\beta}_{yz}$	0.00023	0.00008	0.363	0.278	0.973
	Conditional weighted MICE	$\hat{\beta}_{y0}$	0.00017	0.00002	0.119	0.108	0.989
		$\hat{\beta}_{yx}$	0.00021	0.00006	0.306	0.243	0.976
		$\hat{\beta}_{yz}$	0.00023	0.00006	0.301	0.241	0.977
	Calibrated-δ adjustment MICE	$\hat{\beta}_{0}$	0.00017	0.00001	0.088	0.082	0.992
		$\hat{\beta}_{yx}$	0.00021	0.00006	0.318	0.251	0.976
		$\hat{\beta}_{yz}$	0.00023	0.00007	0.344	0.267	0.974

* Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M3: missingness in x depends on x and in z depends on (z, y); M4: missingness in x depends on (x, y) and in z depends on y; M5: missingness in x depends on (x, y) and in z depends on (z, y).

4.4.3 *Repeated simulations for assessing performance measures*

In this section, performance measures of calibrated-δ adjustment MICE are examined in a repeated multivariate simulation study. Similar to section 3.6.4, repeated simulations are conducted under missingness models M1, M2, and M5 (table 4.2). Prior to the evaluation of calibrated-δ adjustment MICE, marginal weighted MICE is found to be the preferred method under M1 and M2, while conditional weighted MICE produces the least biased parameter estimates under M5.

The single simulation set-up under M1, M2, and M5 described in section 3.6.2 is performed using $S = 1000$ repetitions and $n = 1000$ observations. A smaller sample size is chosen for repeated simulations to reduce processing time, since the calibrated-δ adjustment needs to be re-estimated after every iteration in each imputation of the chained equations, and the interval bisection algorithm can take a relatively long time. As before, in all MI methods, missing values in x and z are imputed using $M = 10$ imputations and $T = 10$ iterations.

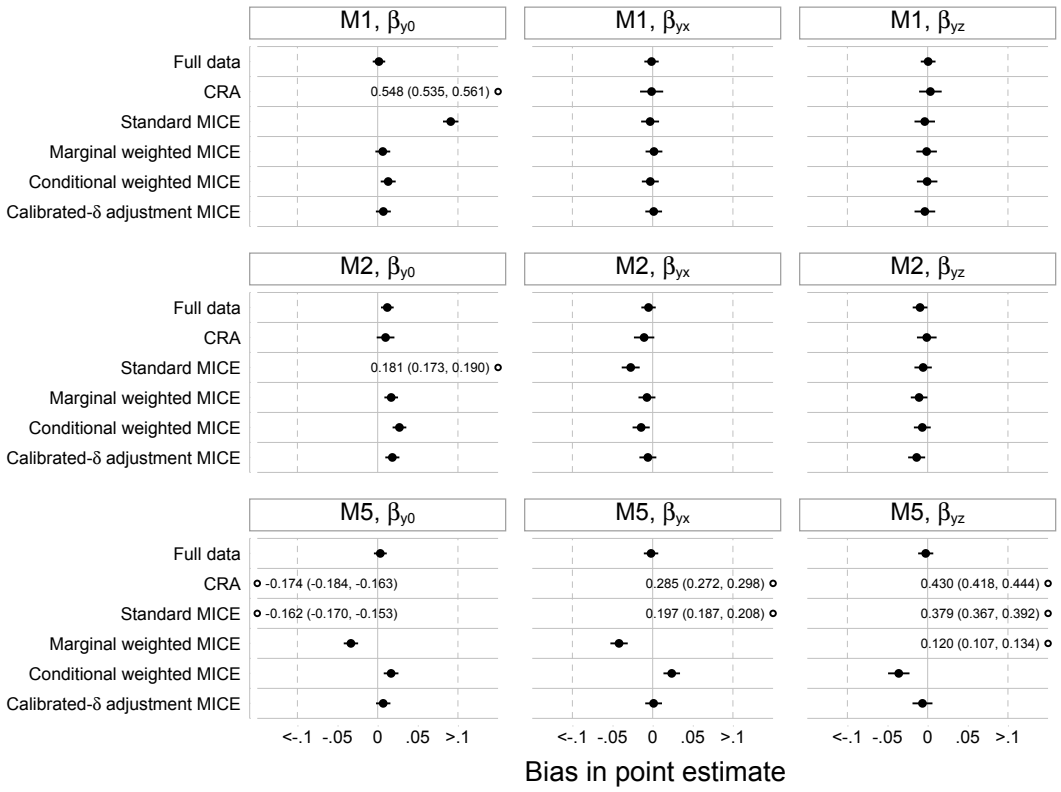
Results are summarised graphically in figures 4.9–4.11. Under M1, when x is MNAR dependent on x and z is MAR conditional on y, calibrated-δ adjustment MICE yields unbiased point estimates and the method is similar to marginal weighted MICE. Empirical standard errors are comparable to the average model standard errors in calibrated-δ adjustment MICE, and they are both similar to that in marginal and conditional weighted MICE. Coverage of 95% CIs attains the nominal level for all three parameters in calibrated-δ adjustment MICE.

Under M2, when x is MNAR dependent on x and z is MNAR dependent on z, there is minimal bias in the estimate of β_{y0} in calibrated-δ adjustment MICE. This bias is similar to the bias seen in marginal weighted MICE and smaller than that in conditional weighted MICE. Calibrated-δ adjustment MICE is unbiased in $\hat{\beta}_{yx}$, while there is negligible bias in $\hat{\beta}_{yz}$.

Under M5, when each of x and y is MNAR dependent on its values and the outcome y, all three parameter estimates are unbiased in calibrated-δ adjustment MICE. Average model standard errors are comparable in calibrated-δ adjustment MICE and conditional weighted MICE. Empirical standard errors are larger than the average model counterparts in calibrated-δ adjustment MICE for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$. This result for the standard errors corresponds to a slight drop in coverage of 95% CIs for these two parameters in calibrated-δ adjustment MICE, with coverage remaining high at around the 92% mark.

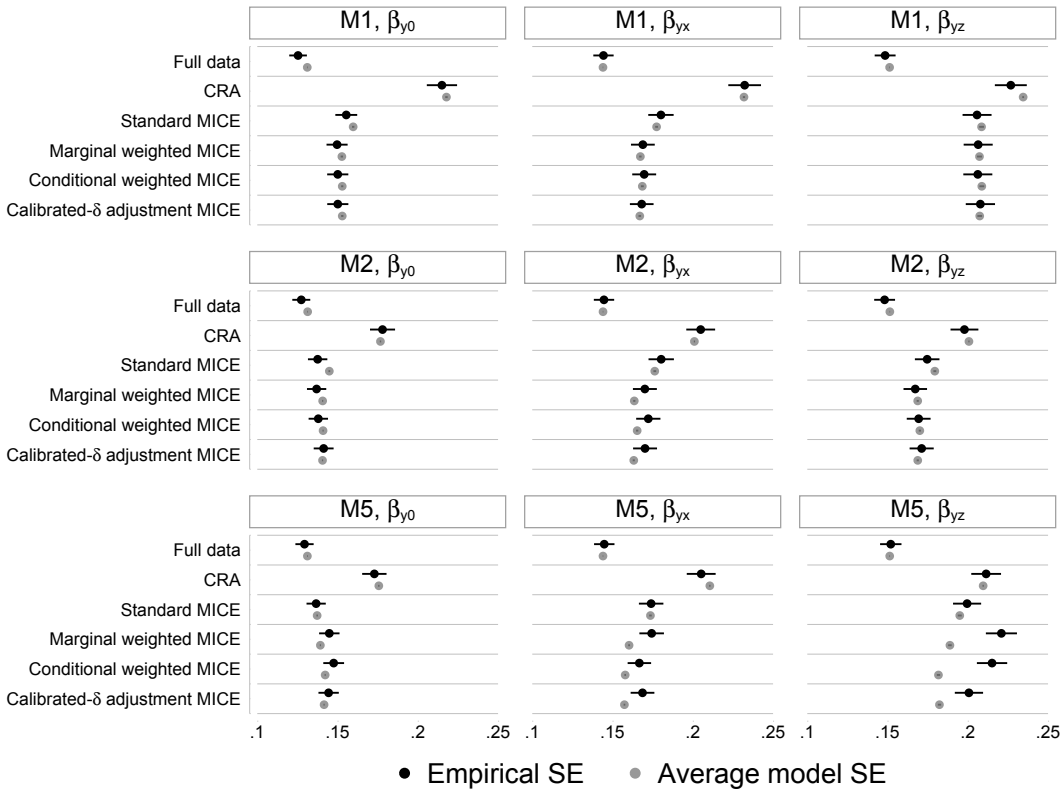
It is also of concern that the chosen full-data sample size for repeated simulations might affect the estimation of the calibrated-δ adjustment by interval bisection. With missing values in both x and z, the observed-data estimation of the log odds ratios in the imputation models for x and z can become less stable. Therefore, the multivariate simulation design is repeated $S = 500$ times with increased sample sizes $n = 3\,000$ and $5\,000$, and the results are presented in appendix B.1. Calibrated-δ adjustment MICE appears unbiased with increased sample sizes. A larger sample size also leads to an overall reduction in the standard errors. Empirical and average model standard errors in calibrated-δ adjustment MICE are more comparable under M5, which improves coverage for β_{yz}.

Figure 4.9. Repeated multivariate simulation study (n = 1 000): bias in point estimates under different missingness mechanisms for x and z.



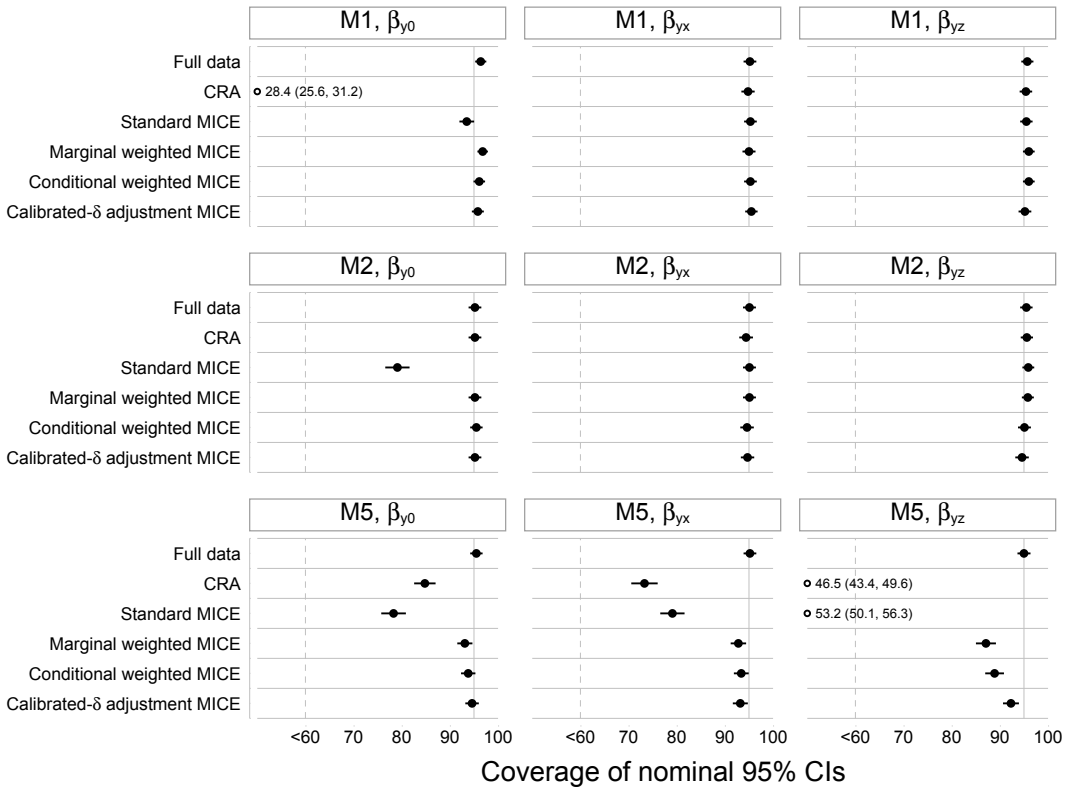
⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); $\beta_0 = 0.5, \beta_x = -1, \beta_z = 1$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure 4.10. Repeated multivariate simulation study (n = 1 000): empirical and average model standard errors under different missingness mechanisms for x and z.



⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 4.11. Repeated multivariate simulation study (n = 1000): coverage of nominal 95% confidence intervals under different missingness mechanisms for x and z.



⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

4.4.4 *Multivariate simulation studies: conclusion and remarks*

Single and repeated multivariate simulation studies with a complete binary outcome variable and two incomplete binary covariates are conducted to explore the extension of univariate calibrated-δ adjustment MI in the MICE algorithm for imputing missing values in more than one covariate.

Marginal weighted MICE and calibrated-δ adjustment MICE yield relatively comparable results when one covariate is MNAR dependent on its values and the other covariate is MAR conditional on the outcome (M1), or both covariates are MNAR dependent on their values (M2). When missingness in each covariate depends on both its values and the outcome (M5), calibrated-δ adjustment MICE appears unbiased with good coverage of at or slightly lower than the 95% level, while bias is seen in both weighted MICE methods. Under M5, the empirical standard errors are larger than the average model standard errors for $\hat{\beta}_{yx}$ and $\hat{\beta}_{yz}$ in calibrated-δ adjustment MICE, and the reason for this mismatch is not clear. This discrepancy between the two standard errors results in a slight decrease in coverage of 95% CIs.

Overall, results from the repeated multivariate simulation study suggest that calibrated-δ adjustment MICE is generally the preferred method under the three missingness mechanisms considered for the incomplete covariates.

4.5 SUMMARY

This chapter proposes and evaluates calibrated-δ adjustment MI as an alternative method to weighting in MI. Calibrated-δ adjustment MI offers a way for incorporating the population-level distribution of the incomplete variable into the imputation process via an offset in the imputation model's intercept. The calibrated-δ adjustment is calculated using the population marginal distribution of the incomplete covariate and its association with other variables in the observed data. Calibrated-δ adjustment MI thus incorporates not only information about the incomplete covariate's population distribution, but also the effects of other variables included in the imputation model for that covariate.

In section 4.2, calibrated-δ adjustment MI is explored analytically in a 2×2 contingency table, with a fully observed binary outcome variable y and a partially observed binary covariate x. It is found that appropriately adjusting the intercept of the imputation model for the incomplete covariate is sufficient to correct bias introduced by missing data under all four missingness mechanisms considered for x. Section 4.3 further evaluates the implementation of calibrated-δ adjustment MI and the method's performance measures in univariate simulation studies of the same setting. Table 4.4 summarises the results of the univariate simulation studies for calibrated-δ adjustment MI in comparison with other methods for handling missing values in the covariate

Table 4.4. Analytic and univariate simulation studies: summary of bias in the analysis model's parameter estimates under different missingness mechanisms for the incomplete covariate x.

Missingness in x depends on	Method for missing data in x	Biased estimation of	
		β_0	β_x
Neither y nor x	CRA	No	No
	Standard MI	No	No
	Marginal weighted MI	No	No
	Conditional weighted MI	No	No
	Calibrated-δ adjustment MI	No	No
y	CRA	Yes	No
	Standard MI	No	No
	Marginal weighted MI	Yes	Yes
	Conditional weighted MI	No	No
	Calibrated-δ adjustment MI	No	No
x	CRA	No	No
	Standard MI	Yes	No
	Marginal weighted MI	No	No
	Conditional weighted MI	Yes	Yes
	Calibrated-δ adjustment MI	No	No
x and y	CRA	Yes	Yes
	Standard MI	Yes	Yes
	Marginal weighted MI	Yes	Yes
	Conditional weighted MI	Yes	Yes
	Calibrated-δ adjustment MI	No	No

* Note: analysis model: $\text{logit}\left[p\left(y = 1 \mid x\right)\right] = \beta_0 + \beta_x x$; y (complete) and x (incomplete) are binary variables, taking values 0 or 1.

x. When missingness in x depends on both the values of x and the outcome y, calibrated-δ adjustment MI removes bias in point estimates that is still present in marginal and conditional weighted MI. This method also has comparable empirical and average model standard errors, and correct coverage of 95% CIs.

The proposed calibrated-δ adjustment MI method for univariate missing data is also adapted for use in the MICE algorithm for imputing missing values in more than one incomplete covariate, accounting for their population marginal distributions. This extension is explored in single and repeated multivariate simulation studies of a three-way contingency table. The analysis model is a logistic regression of a complete binary outcome variable y on two incomplete binary covariates x and z. Simulation results again suggest that under the several missingness mechanisms considered for the incomplete covariates, calibrated-δ adjustment MICE is the preferred method for handling missing values, compared to marginal and conditional weighted MICE and standard MICE.

So far, the development and evaluation of marginal and conditional weighted MI and calibrated-δ adjustment MI focus on the setting where both the outcome variable and the incomplete covariate(s) are binary. The next chapter investigates the application of the weighted MI and calibrated-δ adjustment MI methods in a univariate missing data setting where the fully observed outcome variable is continuous.

5

Population-calibrated multiple imputation of a binary covariate when the outcome variable is continuous

5.1 INTRODUCTION

In chapters 3 and 4, the proposed population-calibrated multiple imputation (MI) methods, including marginal and conditional weighted MI and calibrated-δ adjustment MI, are developed for utilising external information containing the incomplete variable's population-level marginal distribution in the imputation process. Within the context of this thesis, these methods are used for handling missing data in incomplete variables which are included as covariates in the analysis model of interest. The population-calibrated MI methods are explored analytically and via simulation in univariate and multivariate missing data settings where both the outcome variable and the incomplete covariate(s) in the analysis model are binary.

This chapter studies the population-calibrated MI methods in a univariate missing data setting, where the incomplete covariate to be imputed is binary as before but the fully observed outcome variable is continuous. Following the same theme from the previous chapters, the methods are evaluated using progressively increasing realism of the missingness mechanisms for the incomplete covariate.

In section 5.2, marginal and conditional weighted MI and calibrated-δ adjustment MI are evaluated in a univariate simulation study, with a fully observed continuous outcome variable and a partially observed binary covariate. Using repeated simulations, the frequentist properties of the methods are studied and compared to that in standard MI and complete record analysis (CRA). This investigation is conducted to examine whether the population-calibrated MI methods are valid under missingness mechanisms investigated previously when the outcome variable is binary. As shown in this simulation study, when the outcome variable is continuous and missingness in the covariate depends on both its values and the outcome, adjusting the imputation model's intercept based on the population marginal distribution of the incomplete covariate is not sufficient to remove bias introduced by missing data. This finding is due to the presence of a second sensitivity parameter for the covariate–outcome association, which represents how this association differs in the observed and missing data.

Section 5.3 provides a theoretical justification of the additional sensitivity parameter for the covariate–outcome association in the imputation model. A proof-of-concept example based on the set-up of the Heckman model [24] (section 2.5.2) is used to demonstrate that, under a data generating mechanism similar to that used in the above simulation study, the presence of a second sensitivity parameter is expected when the incomplete covariate is missing not at random dependent on its values and the outcome.

Once the presence of the second sensitivity parameter is detected, the problem becomes eliciting the second sensitivity parameter, followed by deriving the calibrated-δ adjustment in the imputation model's intercept. This intercept adjustment is calculated conditional on the incomplete covariate's population marginal distribution and the elicited value of the second sensitivity parameter. This process is demonstrated in a univariate simulation study in section 5.4. In this simulation study, calibrated-δ adjustment MI is evaluated and compared to marginal and conditional weighted MI, standard MI, and CRA when the second sensitivity parameter is fixed to its full-data (i.e. 'correct') estimate.

5.2 UNIVARIATE SIMULATION STUDY

This chapter starts with a univariate simulation study to examine performance measures of marginal and conditional weighted MI and calibrated-δ adjustment MI for handling missing data in an incomplete binary covariate when the complete outcome variable in the analysis model is continuous. The aim of this simulation study is to evaluate the finite-sample properties of marginal and conditional weighted MI and calibrated-δ adjustment MI in this setting. In particular, the properties of interest are bias in parameter estimates, efficiency, and coverage of 95% confidence intervals (CI). This study also aims to examine whether the proposed population-calibrated MI methods are valid under missingness mechanisms considered previously for a binary outcome variable.

5.2.1 Method

The analysis model in this simulation study is a linear regression of a fully observed, normally distributed outcome variable y on an incomplete binary covariate x. As before, marginal and

conditional weighted MI and calibrated-δ adjustment MI are compared to standard MI and CRA under different models of the missingness mechanism for x.

The data generating mechanism and analysis procedures are as follows.

1. Simulate $n = 5\,000$ complete values of the binary covariate x taking values 0 or 1 and the continuous normally distributed outcome y from the following models

$$\begin{aligned} x &\sim \text{Bernoulli}\left(p_x^{\text{pop}} = 0.7\right); \\ y &= \beta_0 + \beta_x x + \varepsilon; \\ \varepsilon &\overset{iid}{\sim} \text{N}\left(0, \sigma^2\right), \end{aligned} \tag{5.1}$$

   where β_0 and β_x are arbitrarily set to -0.5 and 1, respectively. A standard deviation of $\sigma = 0.9$ is chosen to achieve a coefficient of determination $R^2 = 0.2$. The same values of the β coefficients and σ are used throughout to make bias comparable across all simulation settings;
2. Simulate a binary indicator of response r of x from each of the selection models M1–M4 (table 5.1). Under M1–M4, α_y and α_x are set to -1.5 to reflect a strong odds ratio (OR) of observing x (OR = 0.22). For all selection models, α_0 is altered to achieve approximately 45% of missing values in x. The values of $\alpha_0 = 0.25; 0.6; 1.25$; and 1.75 appear to work well for M1–M4, respectively;
3. For $i = 1, \ldots, 5000$, set x_i to missing if $r_i = 0$;
4. Impute missing values in x $M = 50$ times using standard MI, marginal and conditional weighted MI, and calibrated-δ adjustment MI in turn;
5. For each MI method, estimate parameters of the analysis model (5.1) in each completed dataset and combine the results using Rubin's rules [20, 21].

This process is repeated $S = 1\,000$ times under each of the four missingness models M1–M4, so the same set of simulated independent datasets is used to compare the four MI methods under the same missingness scenario, but a different set of datasets is generated for each missingness scenario [77]. The parameters of interest are β_0 and β_x. Bias in the estimates of the β coefficients, efficiency in terms of the empirical standard errors, and coverage of 95% CIs are calculated over 1000 repetitions for each combination of the simulation settings [78], with analyses of the full data and complete records also provided for comparison.

All simulations are performed in Stata 14 [44]; simulated datasets are analysed using the community-contributed command `simsum` [78]. As before, `mi impute logit` [75] is used for standard MI, my command `mi impute wlogit` [73] for marginal and conditional weighted MI, and `mi impute logit, offset` [75] for calibrated-δ adjustment MI.

Table 5.1. Univariate simulation study: models for missingness in x.

Linear predictor of selection model $\text{logit}\left[p\left(r = 1 \mid x, y\right)\right]$	Label
α_0	M1
$\alpha_0 + \alpha_y y$	M2
$\alpha_0 + \alpha_x x$	M3
$\alpha_0 + \alpha_x x + \alpha_y y$	M4

* Note: r: response indicator of x.

In calibrated-δ adjustment MI, the following imputation model

$$\text{logit}\left[p\left(x = 1 \mid y\right)\right] = \theta_0 + \theta_y y$$

is fitted to the complete records, and the estimates of θ_0^{obs} and θ_y^{obs} are obtained. The calibrated-δ adjustment is then calculated based on the assumption that the association between y and x is the same among the observed and missing x, $\theta_y^{\text{obs}} = \theta_y^{\text{mis}}$. Since y is now continuous, this assumption cannot easily be verified in an analytical approach, as in the previous 2×2 contingency table setting (section 4.2.1). However, bias in parameter estimates is expected to be detectable over repeated simulations if this assumption does not hold. Under this assumption, the probability of $x = 1$ among those with missing data in x can therefore be written as

$$p\left(x = 1 \mid r = 0\right) = \frac{1}{n^{\text{mis}}} \sum_{i=1}^{n^{\text{mis}}} \text{expit}\left[\left(\theta_0^{\text{obs}} + \delta\right) + \theta_y^{\text{obs}} y_i\right],$$

and the calibrated-δ adjustment can again be estimated using interval bisection [81, 82] (or any other root-finding method).

5.2.2 *Results*

Results of the univariate simulation study are summarised in figures 5.1–5.3. Full data and CRA again produce results as expected. Point estimates are always unbiased in the full data with the smallest standard errors and correct coverage, except for $\hat{\beta}_0$ under M3 where there is a small over-coverage due to the empirical standard error being slightly smaller than the average model standard error. CRA is unbiased when x is missing completely at random (MCAR, M1) or missing not at random (MNAR) dependent on x (M3). Under these missingness mechanisms, coverage of CRA attains the nominal level. CRA is severely biased under the other two missingness mechanisms, with coverage equal to 0.

Under M1, when x is MCAR, all methods under evaluation are unbiased. Empirical standard errors of marginal and conditional weighted MI and calibrated-δ adjustment MI are slightly higher than the average model standard errors for $\hat{\beta}_x$, leading to very slight under-coverage of 95% CIs.

Under M2, when x is missing at random (MAR) conditional on y, standard MI is, by design, unbiased with correct standard errors and coverage. Conditional weighted MI and calibrated-δ adjustment MI are also unbiased. While the empirical and average model standard errors are similar in conditional weighted MI and coverage of the method is correct, the empirical standard errors are larger than the average model standard errors in calibrated-δ adjustment MI. This increase in the empirical standard errors leads to a small drop in coverage of the method to just below 94%. Marginal weighted MI is biased in both parameter estimates under this missingness mechanism, and coverage decreases substantially to less than 5%.

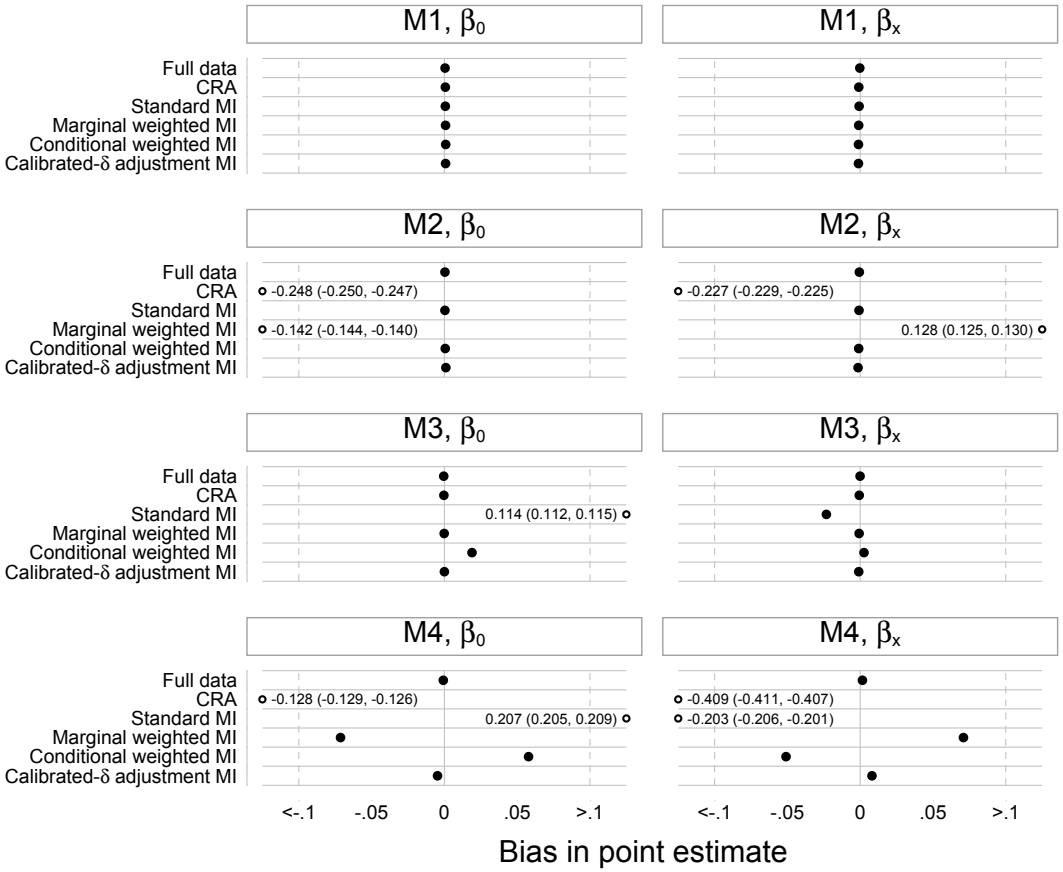
Under M3, when x is MNAR dependent on x, standard MI produces noticeable bias in both point estimates, as expected. Coverage of the method decreases to a larger extent for $\hat{\beta}_0$ compared to $\hat{\beta}_x$, since bias is more severe in $\hat{\beta}_0$. Conditional weighted MI is also biased in both parameter estimates, with larger bias in $\hat{\beta}_0$. Under this missingness mechanism, bias in conditional weighted MI is more noticeable when y is continuous compared to when y is binary. Marginal weighted MI and calibrated-δ adjustment MI are unbiased with similar empirical and average model standard

errors, and coverage of both methods is at the expected 95% level.

Under M4, when x is MNAR dependent on x and y , none of the methods yield unbiased point estimates of β_0 and β_x. However, the magnitude of bias appears to be the smallest in calibrated-δ adjustment MI. Although the empirical standard errors are larger than the average model standard errors in calibrated-δ adjustment MI, coverage remains high, exceeding the 90% level. Due to the substantial bias in point estimates in the other methods, their standard errors are not comparable and coverage is therefore low.

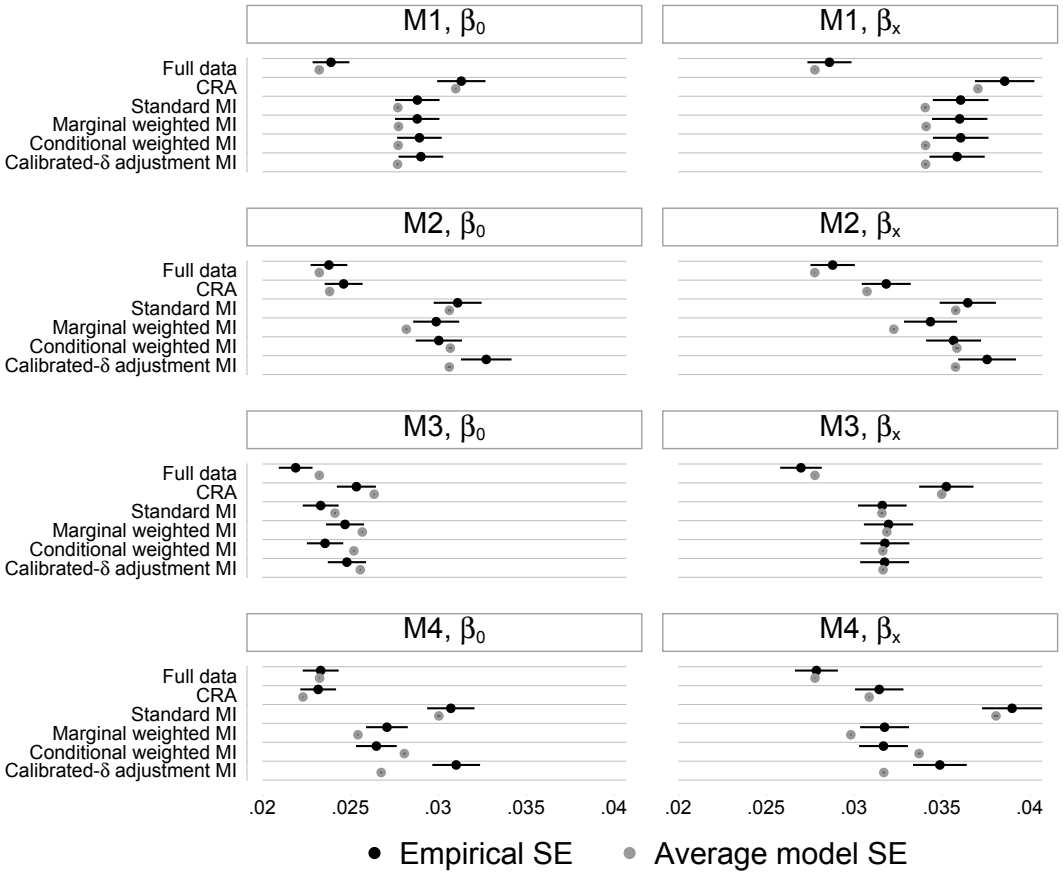
To explore the results for higher and lower degrees of uncertainty in the full data, the univariate simulation study is repeated using the same method (section 5.2.1) but for values of σ = 1.95 and 0.45, which correspond to coefficients of determination R^2 = 0.05 and 0.5, respectively. Results mainly vary with the changes in R^2 under M2 and M4. Overall, there is bias in point estimates in calibrated-δ adjustment MI under M4, but bias decreases with higher R^2. Under M2, although calibrated-δ adjustment MI appears unbiased, the empirical and average model standard errors do not match, and coverage is slightly above or below the nominal level. When R^2 is low, conditional weighted MI is unbiased under M2 and less biased than calibrated-δ adjustment MI under M4, with an over-coverage of 95% CIs under both missingness mechanisms. When R^2 is high, conditional weighted MI is unbiased and achieves the correct coverage under M2; the method is noticeably biased with poor coverage under M4. These results can be found in appendix C.1.

Figure 5.1. Univariate simulation study (R^2 = 0.2): bias in point estimates under different missingness mechanisms for x.



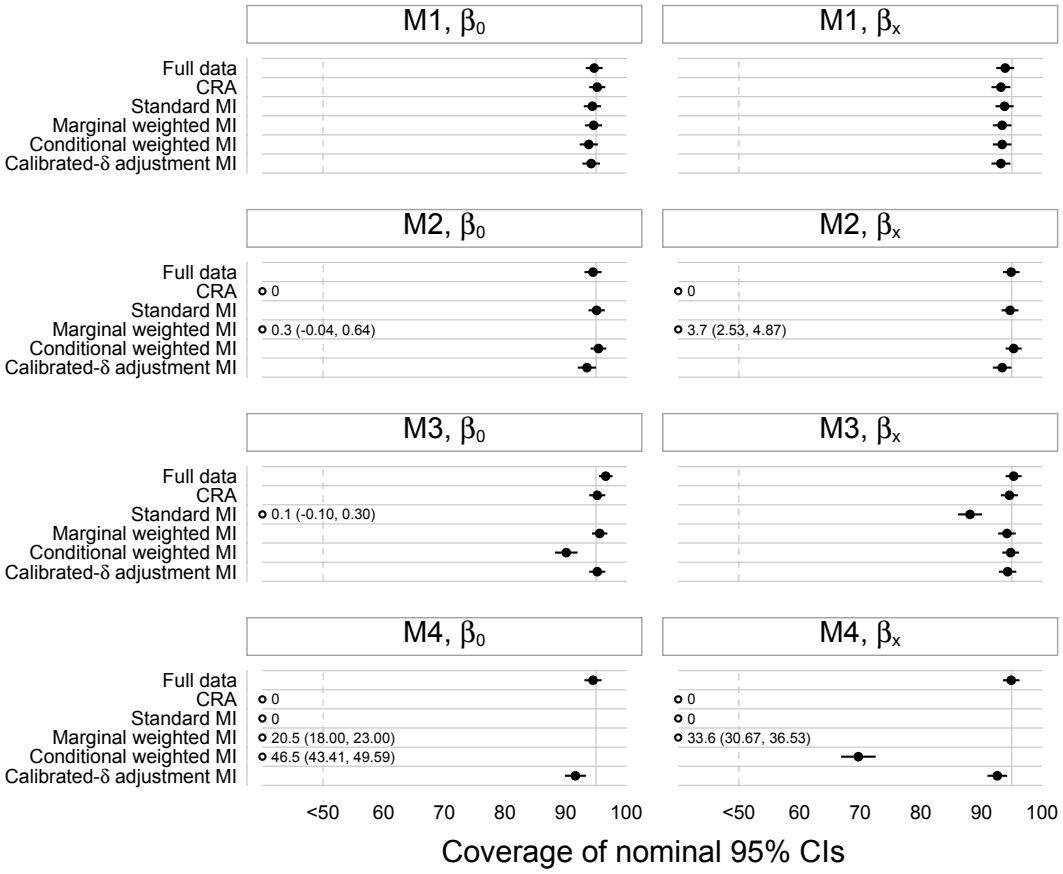
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); $\beta_0 = -0.5, \beta_x = 1$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure 5.2. Univariate simulation study (R^2 = 0.2): empirical and average model standard errors under different missingness mechanisms for x.



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Figure 5.3. Univariate simulation study (R^2 = 0.2): coverage of nominal 95% confidence intervals under different missingness mechanisms for x.



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

5.2.3 *Exploration of the second sensitivity parameter*

Results in the univariate simulation study suggest that the calibrated-δ adjustment MI method produces unbiased parameter estimates when missingness in the incomplete covariate x depends on either the values of x or the outcome variable y. Conversely, the method is valid in all but the last missingness mechanism (M4) considered, under which missingness in x depends on both x and y. Previously, it is noted that the calibrated-δ adjustment MI method is implemented based on the assumption that the association between x and y is the same whether x is observed or missing, i.e. $\theta_y^{\text{obs}} = \theta_y^{\text{mis}}$. Bias in point estimates seen in calibrated-δ adjustment MI under M4 might suggest that, while this assumption might hold under the other missingness mechanisms, it is violated under M4.

To explore whether there is empirical support for this hypothesis, the full datasets in S = 1000 simulation repetitions under each of the four missingness mechanisms are recreated using the random number generator states that correspond to each of the repetitions, and the following model is fitted to each full dataset (i.e. before values in x are set to missing)

$$\text{logit}\left[p\left(x = 1 \mid y, r\right)\right] = \theta_0 + \theta_y y + \theta_r r + \theta_{yr} yr,$$

Table 5.2. Univariate simulation study ($R^2 = 0.2$): mean and standard deviation (SD) of the full-data estimates of θ_r and θ_{yr} over $S = 1000$ simulation repetitions and the number of times each of the null hypotheses $H_0 : \theta_r = 0$ and $H_0 : \theta_{yr} = 0$ is rejected at the 5% level.

Missingness model	$\bar{\hat{\theta}}_r$	$\mathrm{SD}(\hat{\theta}_r)$	$\bar{\hat{\theta}}_{yr}$	$\mathrm{SD}(\hat{\theta}_{yr})$	Number of times $H_0 : \theta_r = 0$ rejected	Number of times $H_0 : \theta_{yr} = 0$ rejected
M1	0.0021	0.0677	−0.0026	0.0912	41	54
M2	0.0029	0.0825	−0.0023	0.1028	47	49
M3	−1.5017	0.0835	−0.0027	0.0966	1000	47
M4	−1.4972	0.0948	0.0934	0.1128	1000	134

* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y).

where r is the response indicator of x. Full-data estimates of θ_r and θ_{yr} are then obtained from the above model. The parameter θ_r of the response indicator of x represents the difference between θ_0^{obs} and θ_0^{mis}, which is the calibrated-δ adjustment in the univariate simulation study and is and now referred to as δ_0. The parameter θ_{yr} of the interaction between the outcome y and the response indicator of x represents the difference between θ_y^{obs} and θ_y^{mis}. This parameter, which is assumed to be 0 in the above simulation study (section 5.2.1), is now referred to as δ_y.

Table 5.2 presents the mean and standard deviation of the estimates of θ_r and θ_{yr} over $S = 1000$ simulation repetitions, which are defined as

$$\bar{\hat{\theta}} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_s;$$

$$\mathrm{SD}(\hat{\theta}) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\theta}_s - \bar{\hat{\theta}}\right)^2},$$

respectively. In addition, the number of times each of the following null hypotheses

$$H_0 : \theta_r = 0, \quad \text{and} \quad H_0 : \theta_{yr} = 0$$

is rejected at the 5% level, where the Wald test p-value is less than 0.05, is also counted and presented in table 5.2.

Under M1 and M2, when x is MCAR or MAR conditional on y, the means of $\hat{\theta}_r$ and $\hat{\theta}_{yr}$ are both relatively close to 0, and both hypotheses are rejected in around 5% of the simulation repetitions. This suggests that no adjustment is needed in either θ_0^{obs} or θ_y^{obs} in the imputation model for x, and standard MI is valid according to the theory of MI. Under M3, when x is MNAR dependent on x, the mean of the θ_r estimates is around −1.5, which is the value of the coefficient α_x used to generate missingness in x; the mean of the estimates of θ_{yr} is again close to 0. In addition, while the second hypothesis regarding θ_{yr} is rejected in about 5% of the simulation repetitions as before, the first hypothesis regarding θ_0 is rejected in all 1000 repetitions. This result is consistent with keeping θ_y^{obs} unadjusted and changing θ_0^{obs} by an adjustment δ_0 in the imputation model for x.

Under M4, when x is MNAR dependent on x and y, the mean of the θ_r estimates is still close to α_x (as it must be when y is binary); however, there is an increase in the magnitude of the mean of the θ_{yr} estimates. The hypothesis regarding θ_{yr} is now rejected in 13% of the simulation repetitions, which suggests that θ_y^{obs} is also different from θ_y^{mis}. The number of times the null

hypothesis $\theta_{yr} = 0$ is rejected increases for lower values of R^2 (appendix C.1). Therefore, in order to correct bias under this missingness mechanism for x, a second adjustment, or *sensitivity parameter*, δ_y is needed in θ_y^{obs} in addition to the existing adjustment δ_0 in θ_0^{obs}.

The presence of the interaction term θ_{yr} between the outcome y and the response indicator r of x in the logistic regression model for x in the full data implies that there is also an interaction term β_{xr} between x and r in the linear regression model for y in the full data, where

$$y = \beta_0 + \beta_x x + \beta_r r + \beta_{xr} xr + \varepsilon;$$
$$\varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

The presence of the interaction term β_{xr} is somewhat unexpected, given that data in y are simulated using model (5.1), which does not include an interaction between x and r. This interaction might be induced by the association between r, x, and y in the missingness model for x. This mechanism is explored in the next section, where the difference between θ_y^{obs} and θ_y^{mis} is demonstrated mathematically in a setting based on the Heckman model [24] (section 2.5.2). This analysis confirms the presence of the second sensitivity parameter when missingness in x depends on x and y.

5.3 Theoretical justification of the additional sensitivity parameter

This section describes a proof-of-concept example which provides a theoretical justification of the interaction between the outcome variable and the missingness indicator of the incomplete covariate in the logistic regression imputation model for the covariate, in the scenario when the covariate is MNAR dependent on its values and the outcome. This supports the empirical findings in section 5.2.2 regarding the presence of a second sensitivity parameter for the covariate–outcome association in the imputation model for the incomplete covariate.

This working example is set up as follows. Let x denote the binary covariate taking values $l = 0$ or 1 where $x \sim \text{Bernoulli}(p_x)$, and y denote the continuous outcome variable which is normally distributed with mean $\beta_0 + \beta_x x$ and variance $\sigma^2_{y|x}$.

Missingness in x is defined through a latent variable z following a normal distribution with mean $\alpha_0 + \alpha_x x + \alpha_y y$ and variance 1. Let a fully observed response indicator r take values 1 if $z \geq 0$ and 0 otherwise. The selection model for x is then defined by two regression models

$$y \mid x \sim \mathrm{N}\left(\beta_0 + \beta_x x, \sigma^2_{y|x}\right); \tag{5.2}$$
$$z \mid y, x \sim \mathrm{N}\left(\alpha_0 + \alpha_x x + \alpha_y y, 1\right), \tag{5.3}$$

where x is missing when $z < 0$ (or $r = 0$). This set-up is similar to the Heckman model [24] discussed in section 2.5.2.

Using the law of total expectation, the conditional expectation of z given x can be written as

$$\begin{aligned}
\mathrm{E}(z \mid x) &= \mathrm{E}(\mathrm{E}(z \mid y, x) \mid x) \\
&= \alpha_0 + \alpha_x x + \mathrm{E}\left(\alpha_y y \mid x\right) \\
&= \alpha_0 + \alpha_x x + \alpha_y(\beta_0 + \beta_x x) \\
&= \left(\alpha_0 + \alpha_y \beta_0\right) + \left(\alpha_x + \alpha_y \beta_x\right) x \\
&= \alpha_0^* + \alpha_x^* x.
\end{aligned}$$

The conditional variance of z given x can be expressed as

$$\begin{aligned}\mathrm{Var}\left(z \mid x\right) &= \mathrm{E}\left(\mathrm{Var}\left(z \mid y, x\right) \mid x\right) + \mathrm{Var}\left(\mathrm{E}\left(z \mid y, x\right) \mid x\right) \\ &= 1 + \mathrm{Var}\left(\alpha_0 + \alpha_x x + \alpha_y y \mid x\right) \\ &= 1 + \alpha_y^2 \mathrm{Var}\left(y \mid x\right) \\ &= 1 + \alpha_y^2 \sigma_{y|x}^2 \\ &= \sigma_{z|x}^2.\end{aligned}$$

$f\left(y, z \mid x\right)$ thus jointly follows a bivariate normal distribution, whose density function is

$$\begin{pmatrix} y \\ z \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mu_{y|x} = \beta_0 + \beta_x x \\ \mu_{z|x} = \alpha_0^* + \alpha_x^* x \end{pmatrix} \quad , \quad \begin{pmatrix} \sigma_{y|x}^2 & \rho\sigma_{y|x}\sigma_{z|x} \\ \rho\sigma_{y|x}\sigma_{z|x} & \sigma_{z|x}^2 \end{pmatrix}\right], \tag{5.4}$$

where ρ is the correlation coefficient between y and z, $\rho = \mathrm{corr}\left(y, z \mid x\right)$.

The conditional distribution of z given y and x can then be written as

$$z \mid y, x \sim \mathrm{N}\left(\mu_{z|x} + \rho\frac{\sigma_{z|x}}{\sigma_{y|x}}\left(y - \mu_{y|x}\right), \sigma_{z|x}^2\left(1 - \rho^2\right)\right). \tag{5.5}$$

(5.3) and (5.5) imply that

$$1 = \sigma_{z|x}^2\left(1 - \rho^2\right),$$

from which the correlation ρ between y and z can be derived as

$$\begin{aligned}\rho &= \sqrt{\frac{\sigma_{z|x}^2 - 1}{\sigma_{z|x}^2}} \\ &= \frac{\alpha_y \sigma_{y|x}}{\sqrt{1 + \alpha_y^2 \sigma_{y|x}^2}}.\end{aligned} \tag{5.6}$$

The probability of $x = 1$ conditional on y in the observed data is given by

$$\begin{aligned}p\left(x = 1 \mid y, z \geq 0\right) &= \frac{f\left(z \geq 0 \mid y, x = 1\right) f\left(y \mid x = 1\right) p\left(x = 1\right)}{\sum_{l=0}^{1} f\left(z \geq 0 \mid y, x = l\right) f\left(y \mid x = l\right) p\left(x = l\right)} \\ &= \frac{1}{1 + \dfrac{f\left(z \geq 0 \mid y, x = 0\right) f\left(y \mid x = 0\right) p\left(x = 0\right)}{f\left(z \geq 0 \mid y, x = 1\right) f\left(y \mid x = 1\right) p\left(x = 1\right)}},\end{aligned} \tag{5.7}$$

and, similarly, in the missing data

$$p\left(x = 1 \mid y, z < 0\right) = \frac{1}{1 + \dfrac{f\left(z < 0 \mid y, x = 0\right) f\left(y \mid x = 0\right) p\left(x = 0\right)}{f\left(z < 0 \mid y, x = 1\right) f\left(y \mid x = 1\right) p\left(x = 1\right)}}.$$

From (5.2), the density function of y conditional on x can be expressed as

$$f\left(y \mid x = 1\right) = \frac{1}{\sqrt{2\pi\sigma_{y|x}^2}} \exp\left[-\frac{\left(y - \beta_0 - \beta_x x\right)^2}{2\sigma_{y|x}^2}\right]. \tag{5.8}$$

From (5.5) and (5.6), the truncated density function of $z \geq 0$ given y and x can be written as

$$f\left(z \geq 0 \mid y, x\right) = \Phi\left[\frac{\mu_{z|x} + \rho\frac{\sigma_{z|x}}{\sigma_{y|x}}\left(y - \mu_{y|x}\right)}{\sqrt{1}}\right]$$

$$
\begin{aligned}
&= \Phi\left[\left(\alpha_0^* + \alpha_x^* x\right)\rho\frac{\sigma_{z|x}}{\sigma_{y|x}}\left(y - \beta_0 - \beta_x x\right)\right] \\
&= \Phi\left[\left(\alpha_0 + \alpha_y\beta_0\right) + \left(\alpha_x + \alpha_y\beta_x\right)x + \alpha_y\left(y - \beta_0 - \beta_x x\right)\right] \\
&= \Phi\left[\alpha_0 + \alpha_x x + \alpha_y y\right]. \qquad (5.9)
\end{aligned}
$$

The logistic regression of x conditional on y in the observed data is

$$\operatorname{logit}\left[p\left(x = 1 \mid y, z \geq 0\right)\right] = \theta_0^{\text{obs}} + \theta_y^{\text{obs}} y,$$

from which the probability of $x = 1$ given y in the observed data is given by

$$p\left(x = 1 \mid y, z \geq 0\right) = \operatorname{expit}\left(\theta_0^{\text{obs}} + \theta_y^{\text{obs}} y\right). \qquad (5.10)$$

From (5.7)–(5.10), the linear predictor in the logistic regression of x conditional on y in the observed data can be written as

$$
\begin{aligned}
\theta_0^{\text{obs}} + \theta_y^{\text{obs}} y &= \ln\left[\frac{f\left(z \geq 0 \mid y, x = 1\right) f\left(y \mid x = 1\right) p\left(x = 1\right)}{f\left(z \geq 0 \mid y, x = 0\right) f\left(y \mid x = 0\right) p\left(x = 0\right)}\right] \\
&= \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x + \alpha_y y\right)}{\Phi\left(\alpha_0 + \alpha_y y\right)}\right] + \ln\left\{\frac{\exp\left[-\frac{\left(y - \beta_0 - \beta_x\right)^2}{2\sigma_{y|x}^2}\right]}{\exp\left[-\frac{\left(y - \beta_0\right)^2}{2\sigma_{y|x}^2}\right]}\right\} + \ln\left(\frac{p_x}{1 - p_x}\right) \\
&= \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x + \alpha_y y\right)}{\Phi\left(\alpha_0 + \alpha_y y\right)}\right] - \frac{\left(y - \beta_0 - \beta_x\right)^2}{2\sigma_{y|x}^2} + \frac{\left(y - \beta_0\right)^2}{2\sigma_{y|x}^2} + \ln\left(\frac{p_x}{1 - p_x}\right). \qquad (5.11)
\end{aligned}
$$

When $y = 0$, (5.11) becomes

$$\theta_0^{\text{obs}} = \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x\right)}{\Phi\left(\alpha_0\right)}\right] - \frac{2\beta_0\beta_x + \beta_x^2}{2\sigma_{y|x}^2} + \ln\left(\frac{p_x}{1 - p_x}\right), \qquad (5.12)$$

and similarly, when $y = 1$

$$\theta_0^{\text{obs}} + \theta_y^{\text{obs}} = \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x + \alpha_y\right)}{\Phi\left(\alpha_0 + \alpha_y\right)}\right] - \frac{\beta_x^2 + 2\beta_0\beta_x - 2\beta_x}{2\sigma_{y|x}^2} + \ln\left(\frac{p_x}{1 - p_x}\right). \qquad (5.13)$$

The log odds ratio of x for y in the observed data can be derived from (5.12) and (5.13) as

$$\theta_y^{\text{obs}} = \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x + \alpha_y\right)\Phi\left(\alpha_0\right)}{\Phi\left(\alpha_0 + \alpha_y\right)\Phi\left(\alpha_0 + \alpha_x\right)}\right] + \frac{\beta_x}{\sigma_{y|x}^2}. \qquad (5.14)$$

In the missing data, the logistic regression of x conditional on y is

$$\operatorname{logit}\left(x = 1 \mid y, z < 0\right) = \theta_0^{\text{mis}} + \theta_y^{\text{mis}} y.$$

Following the same steps as above, the log odds and log odds ratio of x for y in the missing data can be written as

$$\theta_0^{\text{mis}} = \ln\left[\frac{1 - \Phi\left(\alpha_0 + \alpha_x\right)}{1 - \Phi\left(\alpha_0\right)}\right] - \frac{2\beta_0\beta_x + \beta_x^2}{2\sigma_{y|x}^2} + \ln\left(\frac{p_x}{1 - p_x}\right); \qquad (5.15)$$

$$\theta_y^{\text{mis}} = \ln\left\{\frac{\left[1 - \Phi\left(\alpha_0 + \alpha_x + \alpha_y\right)\right]\left[1 - \Phi\left(\alpha_0\right)\right]}{\left[1 - \Phi\left(\alpha_0 + \alpha_y\right)\right]\left[1 - \Phi\left(\alpha_0 + \alpha_x\right)\right]}\right\} + \frac{\beta_x}{\sigma_{y|x}^2}. \qquad (5.16)$$

From (5.12) and (5.15), the difference between the log odds of x in the observed and missing data is given by

$$\theta_0^{\text{mis}} - \theta_0^{\text{obs}} = \ln\left[\frac{1 - \Phi\left(\alpha_0 + \alpha_x\right)}{1 - \Phi\left(\alpha_0\right)} \cdot \frac{\Phi\left(\alpha_0\right)}{\Phi\left(\alpha_0 + \alpha_x\right)}\right]. \qquad (5.17)$$

This difference is a function of the parameters defining the latent variable z which governs missingness in x.

Similarly, (5.14) and (5.16) show that there is also a difference between the log odds ratios of x for y in the observed and missing data. This difference is again a function of the parameters used for generating missingness in x, and can be written as

$$\theta_y^{\text{mis}} - \theta_y^{\text{obs}} = \ln\left\{\frac{\left[1-\Phi\left(\alpha_0+\alpha_x+\alpha_y\right)\right]\left[1-\Phi\left(\alpha_0\right)\right]}{\left[1-\Phi\left(\alpha_0+\alpha_y\right)\right]\left[1-\Phi\left(\alpha_0+\alpha_x\right)\right]}\cdot\frac{\Phi\left(\alpha_0+\alpha_y\right)\Phi\left(\alpha_0+\alpha_x\right)}{\Phi\left(\alpha_0+\alpha_x+\alpha_y\right)\Phi\left(\alpha_0\right)}\right\}. \tag{5.18}$$

When x is MCAR, $\alpha_x = \alpha_y = 0$, and it follows that the differences in the log odds and log odds ratios in (5.17) and (5.18) are both equal to 0. When x is MAR conditional on y, $\alpha_x = 0$, and again both (5.17) and (5.18) are equal to 0. This implies that under these two missingness mechanisms, no adjustment is needed in the imputation model, and standard MI is the valid approach as the theory suggests. In contrast, when x is MNAR dependent on x, $\alpha_y = 0$, which means that the difference in the log odds ratios in (5.18) is 0 while (5.17) is not equal to 0, so the log odds in the observed and missing data are different. This is consistent with findings in the univariate simulation study in section 5.2, which suggest that adjusting the imputation model's intercept is sufficient to remove bias when missingness in x depends on x. Lastly, when missingness in x depends on x and y, neither α_x nor α_y is 0, and thus both (5.17) and (5.18) are not equal to 0. This finding confirms the presence of the second sensitivity parameter for the association between x and y in the imputation model for x, and further demonstrates why adjusting only the intercept of the imputation model is not sufficient to remove bias introduced by the MNAR mechanism in this case. Nevertheless, using the intercept adjustment when x is MNAR dependent on both x and y is expected to yield little bias when α_y is small compared to the other selection parameters.

These calculations are verified by simulating a large full dataset of size $n = 1\,000\,000$ for x, y, and z using pre-defined values of the β and α coefficients and $\sigma_{y|x}$ (appendix C.2). Analytic results obtained from following the above calculation steps are compared to the empirical results obtained from fitting the relevant models to the simulated dataset, in order to verify consistency.

As expected, this working example demonstrates the presence of a sensitivity parameter for θ_0^{obs}. In addition, the calculations (and the empirical findings) in this example also confirm the presence of an additional sensitivity parameter for θ_y^{obs}, as seen in the univariate simulation study in sections 5.2.2 and 5.2.3.

5.4 Univariate simulation study: when the second sensitivity parameter is fixed to its full-data estimate

Findings in the univariate simulation study and calculations in sections 5.2 and 5.3 explain why adjusting the intercept θ_0^{obs} of the imputation model for x alone cannot sufficiently account for bias introduced by the inclusion of y in the MNAR mechanism (M4) for x. This deficiency indicates that under this missingness mechanism, knowing the population-level marginal distribution of the covariate is not enough to correctly recover the second sensitivity parameter for the association between the covariate and the outcome variable in the missing data.

The problem then becomes exploring the sensitivity of inference for a range of values of the

second sensitivity parameter δ_y. This evaluation can be conducted by eliciting δ_y and using the population distribution of the incomplete covariate x to derive δ_0, given each elicited value of δ_y. The estimate of interest and its 95% CIs can be graphed against the various choices of δ_y.

The following sections revisit the univariate simulation study presented in section 5.2, in order to explore bias in calibrated-δ adjustment MI under missingness mechanism M4, when the sensitivity parameter for θ_y^{obs} is fixed to its estimate obtained in full data.

5.4.1 *Method*

Based on the findings highlighted above, this section focuses on missingness mechanism M4 (table 5.1), under which missingness in the incomplete covariate x depends on both the values of x and the outcome variable y.

The data generating mechanism and simulation procedures follow the method described in section 5.2.1. For each of the $S = 1\,000$ simulation repetitions, the same repetition-wise state of the random number generator used previously is set to recreate the full dataset and missing values in x, in order to make bias comparable to previous simulations. In the presence of the second sensitivity parameter, the logistic regression model for x conditional on y in the full data is given by

$$\text{logit}\left[p\left(x = 1 \mid y, r\right)\right] = \theta_0 + \theta_y y + \theta_r\left(1 - r\right) + \theta_{yr} y\left(1 - r\right). \tag{5.19}$$

This parameterisation implies that the imputation model for x in the missing data can be written as

$$\text{logit}\left[p\left(x = 1 \mid y, r = 0\right)\right] = \left(\theta_0^{\text{obs}} + \delta_0\right) + \left(\theta_y^{\text{obs}} + \delta_y\right) y, \tag{5.20}$$

where δ_0 and δ_y represent the adjustments in the parameter estimates obtained from fitting the logistic regression model for x conditional on y in subjects with observed x.

Since a full dataset is simulated in each simulation repetition, model 5.19 can be fitted to the full data before any values of x are set to missing, and $\hat{\theta}_{yr}$ representing the full-data (i.e. 'correct') value of the adjustment for θ_y is recorded. The probability of $x = 1$ among the missing x can be written as

$$p\left(x = 1 \mid r = 0\right) = \frac{1}{n^{\text{mis}}} \sum_{i=1}^{n^{\text{mis}}} \text{expit}\left[\left(\theta_0^{\text{obs}} + \delta_0\right) + \left(\theta_y^{\text{obs}} + \delta_y\right) y_i\right]. \tag{5.21}$$

Hence, δ_0 can be derived using interval bisection [81, 82] (or any other root-finding method), after δ_y is fixed to $\hat{\theta}_{yr}$ which is estimated in full data in the previous step.

In Stata, this process can be integrated in the imputation via the `offset` option which is specific to `mi impute logit` [75], as before. First, model (5.19) is fitted to the full data, and the estimate of θ_{yr} is stored in a local macro `delta_y`. Next, δ_0 is estimated using interval bisection given the chosen value of δ_y from the previous step, and the estimate is also stored in a local macro `delta_0`. Finally, a variable `offsetvar` containing the offset is created, and MI is performed using the following commands.

```
. generate offsetvar = -(‘delta_0’ + ‘delta_y’*y) * r
. mi impute logit x y, offset(offsetvar) add(50)
```

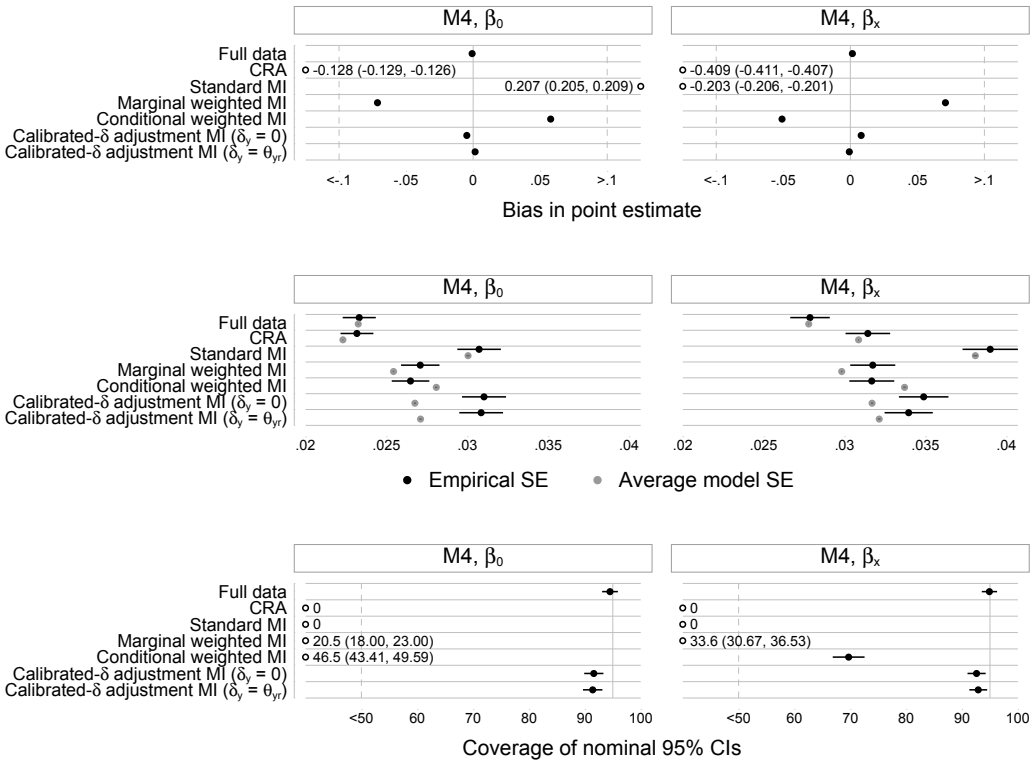
As before, missing values in x are imputed $M = 50$ times, and the parameters of interest, β_0 and β_x, are estimated in each completed dataset and the results are combined using Rubin's rules

[20, 21]. Bias in the β coefficient estimates, efficiency in terms of the empirical standard errors, and coverage of 95% CIs are calculated over 1000 simulation repetitions for each MI method [78], with analyses of the full data and complete records also provided for reference. Stata 14 [44] is used for all aspects of this simulation study.

5.4.2 *Results*

Figure 5.4 presents the results of the univariate simulation study under M4, when x is MNAR conditional on x and y. Results for the full data, CRA, standard MI, marginal and conditional weighted MI, and calibrated-δ adjustment MI assuming $\delta_y = 0$ are discussed in section 5.2.2 and included here for reference.

Figure 5.4. Univariate simulation study ($R^2 = 0.2$): bias in point estimates, empirical and average model standard errors, and coverage of nominal 95% confidence intervals when missingness in x depends on x and y (M4).



* Note: $\beta_0 = -0.5, \beta_x = 1$; error bars are ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

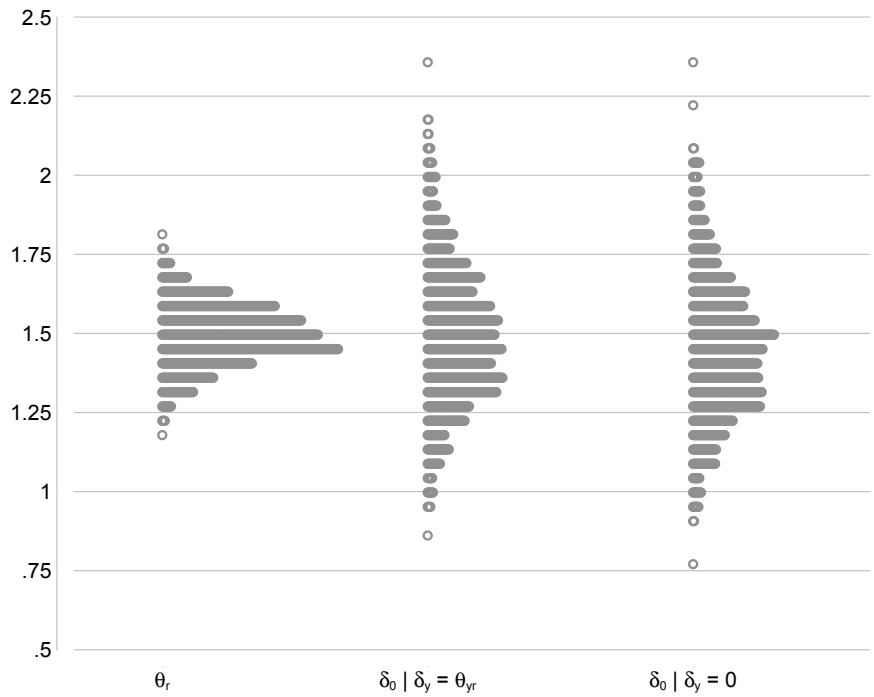
Bias previously seen in calibrated-δ adjustment MI when δ_y is assumed to be 0 is now corrected under calibrated-δ adjustment MI with δ_y fixed to its estimate obtained in full data. This result is as expected, because once the δ_y adjustment is set to the correct value, the remaining sensitivity parameter is δ_0, which can be recovered using the population marginal distribution of the incomplete variable x.

Results regarding the standard errors and coverage are generally similar for the two calibrated-δ adjustment MI approaches (assuming $\delta_y = 0$ and setting δ_y to its full-data estimate). The discrepancy between the empirical and average model standard errors seen in calibrated-δ adjustment MI assuming $\delta_y = 0$ also appears in calibrated-δ adjustment MI with δ_y fixed to its full-data estimate. This dissimilarity between the two standard errors leads to an under-coverage of 95% CIs, with coverage of both parameter estimates remaining above the 90% level. However, this dissimilarity is smaller, particularly for $\hat{\beta}_x$, when δ_y is fixed to the correct value, which results in a slight improvement in coverage for this parameter. Results for $R^2 = 0.05$ and 0.5 are similar and are presented in appendix C.3.

Figure 5.5 shows a comparison of the distributions of $\hat{\theta}_r$ estimated in full data, δ_0 derived assuming δ_y is fixed to its full-data estimate $\hat{\theta}_{yr}$, and δ_0 derived assuming δ_y is equal to 0. Dot-plots of $\hat{\theta}_r$ and δ_0 under two different assumptions for δ_y over $S = 1000$ simulation repetitions are presented, ignoring the sign of the quantities.

While the location of the distributions is comparable for $\hat{\theta}_r$ and δ_0 given $\delta_y = \hat{\theta}_{yr}$ (sample mean = 1.497 and 1.491 over 1000 simulation repetitions, respectively), there is a shift towards 0

Figure 5.5. Univariate simulation study ($R^2 = 0.2$): comparison of $\hat{\theta}_r$ estimated in the full data; calibrated δ_0 derived assuming $\delta_y = \hat{\theta}_{yr}$, where $\hat{\theta}_{yr}$ is estimated in the full data; and calibrated δ_0 derived assuming $\delta_y = 0$ over $S = 1\,000$ simulation repetitions, when missingness in x depends on x and y (M4).



in the distribution of δ_0 assuming $\delta_y = 0$ (sample mean = 1.451 over 1000 simulation repetitions). The assumption of $\delta_y = 0$ therefore leads to an underestimation of the first sensitivity parameter δ_0 under the missingness mechanism M4 investigated in the simulation study.

The spread of the distribution of $\hat{\theta}_r$ is narrower than that of δ_0 given $\delta_y = \hat{\theta}_{yr}$ (SD = 0.095 and 0.218, respectively), while the spread of the distribution is similar for δ_0 given $\delta_y = \hat{\theta}_{yr}$ and δ_0 given $\delta_y = \hat{\theta}_0$ (SD = 0.218 and 0.221, respectively).

This comparison is repeated for $R^2 = 0.05$ and 0.5. The sample mean is less comparable for the three quantities with decreasing values of R^2, and the distributions of δ_0 have wider spreads for higher values of R^2. These results are presented in appendix C.3.

5.4.3 *Univariate simulation studies: conclusion and remarks*

In the first univariate simulation study, the analysis model is a linear regression of a fully observed continuous and normally distributed outcome variable and an incomplete binary covariate. Calibrated-δ adjustment MI works well in terms of bias when the covariate is MCAR, MAR dependent on the outcome, or MNAR dependent on its values.

When the covariate is MNAR conditional on both its values and the outcome, calibrated-δ adjustment MI is biased, and the extent of bias increases with higher variation σ (i.e. lower R^2) in the full data. Under this missingness mechanism, the univariate simulation study and calculations exploring calibrated-δ adjustment MI in sections 5.2 and 5.3 confirm the presence of a second sensitivity parameter for the association between the incomplete covariate and the outcome in the missing data. This second sensitivity parameter does not appear when the outcome variable is binary (sections 4.2.1 and 4.3.2).

Another univariate simulation study is conducted, in which the second sensitivity parameter is fixed to its full-data estimate and the adjustment in the imputation model's intercept is derived using calibrated δ-adjustment MI as before. It is found that this approach can correct bias in parameter estimates which is introduced by the inclusion of the outcome variable in the MNAR model for the covariate. However, the empirical standard errors of the method do not match the average model standard errors, which lead to coverage being slightly over or under the 95% level, depending on the values of R^2 and σ used to generate the full data. The reason for this discrepancy in the standard errors is not clear.

5.5 SUMMARY

This chapter explores the application of marginal and conditional weighted MI and calibrated-δ adjustment MI in a univariate missing data setting, where the incomplete covariate is binary as considered previously, but the outcome variable is now continuous.

As before, four increasingly complex models of the missingness mechanism for the incomplete covariate are examined in a univariate simulation study (section 5.2). Under the first three missingness mechanisms, results for marginal and conditional weighted MI and calibrated-δ adjustment MI are shown to be similar to the setting where the outcome variable is binary. However, calibrated-δ adjustment MI leads to bias in point estimates under the last missingness mechanism when the covariate is MNAR dependent on its values and the outcome. Due to the

inclusion of the continuous outcome variable in the MNAR selection model for the covariate, the assumption of the covariate–outcome association being the same in the observed and missing data is therefore violated. This violation implies that under this missingness mechanism, simply adjusting the imputation model's intercept might not be enough to remove bias caused by missing data in the covariate.

In section 5.3, a working example based on the design of the Heckman model [24] is introduced to confirm the presence of a second sensitivity parameter for the association between the incomplete covariate and the outcome in the imputation model for the covariate.

Further repeated simulations (section 5.4) show that setting the second sensitivity parameter to the correct value and deriving the calibrated-δ adjustment in the imputation model's intercept can correct bias previously seen in calibrated-δ adjustment MI, when the method is implemented with only one sensitivity parameter. In practice, since the full data are not available, the robustness of inference can be explored for several values of the second sensitivity parameter, in which each of the intercept adjustments is derived using the incomplete variable's population distribution.

For a continuous outcome variable, when the calibrated-δ adjustment MI method produces unbiased estimates, it is still not clear why there is a discrepancy between the method's empirical and average model standard errors. Since this discrepancy can subsequently affect coverage of 95% CIs, further investigations are thus required to gain a better understanding of this issue in calibrated-δ adjustment MI when the outcome variable is continuous.

In the next chapter, the methodology of the population-calibrated MI methods developed thus far is applied in two case studies using data from a large UK primary care electronic health record database.

6

Case studies using UK primary care electronic health records

6.1 INTRODUCTION

In previous chapters, the development and implementation of the population-calibrated multiple imputation (MI) methods, including marginal and conditional weighted MI and calibrated δ-adjustment MI, are explored and evaluated in analytic and simulation studies with univariate and then multivariate missing data. In this chapter, the application of these methods is demonstrated using real-life data from a large UK primary care electronic health record database.

Section 6.2 provides an overview of primary care electronic health record databases in the UK, including The Health Improvement Network (THIN) database which is the main data source for the two case studies presented in this chapter. This is followed by a description of how data

are typically recorded in the primary care setting, which gives rise to the problem of missing data in research using primary care databases. Since the methodological development of the population-calibrated MI methods in this thesis is motivated by the incompleteness of ethnicity information in primary care databases, the recording of ethnicity information in primary care is discussed in section 6.3.

Two case studies are presented in sections 6.4 and 6.5, using THIN data to illustrate the application of marginal and conditional weighted MI and calibrated-δ adjustment MI for handling missing values in ethnicity. These methods are also compared to standard MI as well as other simple approaches to missing data. In the first case study, the plausibility of the missing at random assumption for ethnicity is assessed by handling missing values in ethnicity using different methods, estimating the resulting marginal distribution of ethnicity in each method, and comparing that to the corresponding population distribution in the census data. The second case study aims to demonstrate the use of the population-calibrated MI methods for handling missing ethnicity data, when ethnicity is included as a covariate in an analysis model to examine the association between ethnicity and the prevalence of type 2 diabetes diagnoses in primary care.

6.2 UK primary care databases and the issue of missing data

An electronic health record is a digital version of an individual's medical record. Primary care electronic health records refer to the systematic collection of patient information during routine consultations with a general practictioner (GP) or other healthcare professionals in the primary care setting.

Almost the entire UK population is registered with a GP, and under the National Health Service (NHS) the majority of treatments including visits to the GPs are free of charge. GPs act as the gatekeepers of the NHS; they provide the first point of contact for any non-emergency health-related issues, which may then be managed within primary care or referred to secondary care if necessary. UK general practices now have computerised health records, where patient data collected in primary care are routinely documented onto computers by practice staff using a unique patient NHS number. Most information recorded in secondary care including key diagnoses is also fed back to the GPs and added to the patient records.

Several large and well established primary care databases are available in the UK, including the Clinical Practice Research Datalink (CPRD, formerly known as the General Practice Research Database, GPRD) [83], The Health Improvement Network (THIN) [84], and QRESEARCH [85]. These databases provide access to patient-centred health information collected at irregular time points from when the patients first register with their general practices to the time they die or leave the practices, forming longitudinal health records. Data in primary care databases typically include patient demographics (e.g. year of birth, sex, social deprivation), medical records (e.g. symptoms, diagnoses, and referrals to secondary care), prescription information, laboratory test results, lifestyle-related factors (e.g. smoking and alcohol consumption), and measurements of health indicators taken during consultation (e.g. height, body weight, blood pressure, and cholesterol level).

The next section provides background information on The Health Improvement Network

database which is used as the main data source for case studies in this chapter.

6.2.1 *The Health Improvement Network database*

The Health Improvement Network (THIN) primary care database [84] represents a collaboration between In Practice Systems who developed the Vision software used by GPs in the UK to record and manage patient data, and IMS Health who then provide access to the data for use in health research. Data are collected during routine consultations in primary care and regularly downloaded to the database. THIN data used in research are pseudonymised, i.e. the data do not contain patient-identifiable information such as name, exact address or postcode, exact date of birth, or NHS number. Data collection was ethically approved for use in scientific research by the NHS South-East Multicentre Research Ethics Committee. Studies using THIN data must undergo scientific review provided by an independent Scientific Review Committee to help ensure appropriate analysis and interpretation of the data.

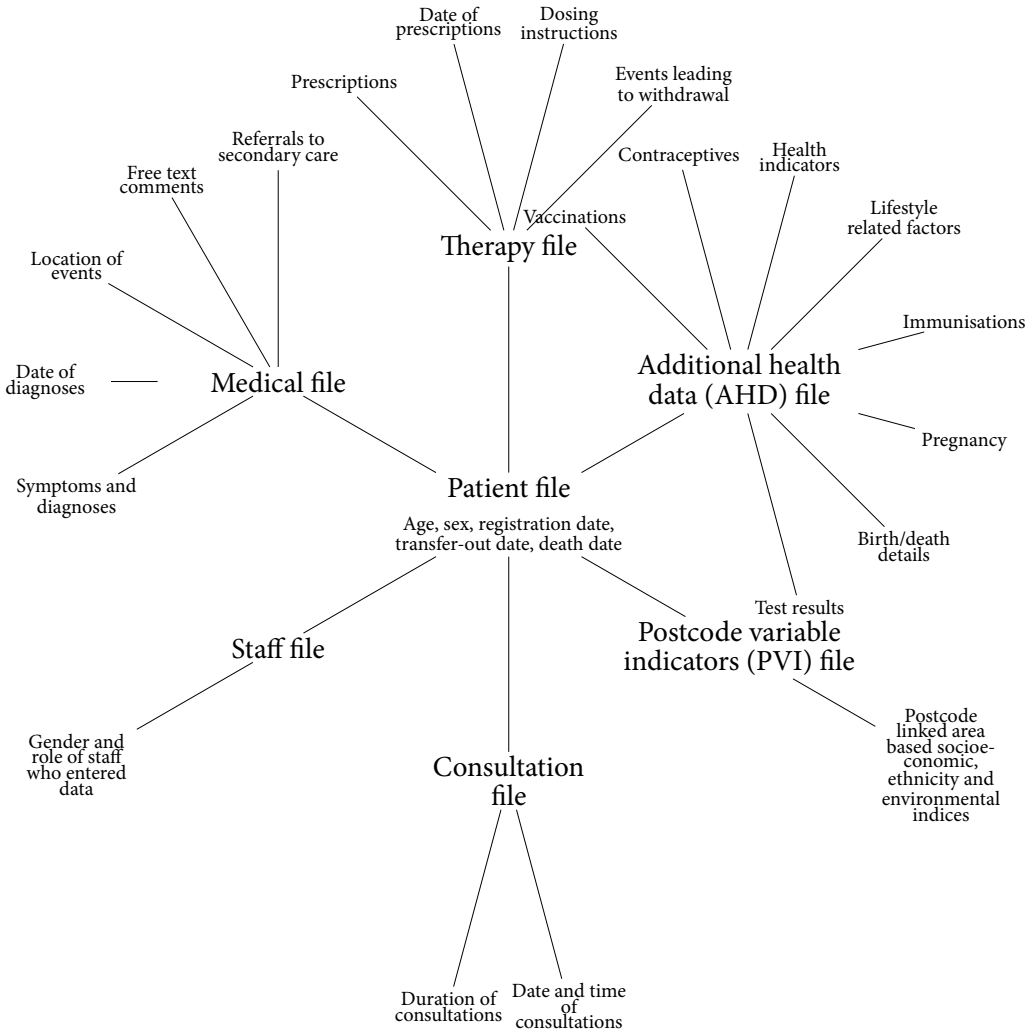
THIN data collection commenced in 2002, but data for some practices date back to the early 1990s. In 2013, the database captured data contributed by more than 12 million patients from 587 general practices in the UK, covering approximately 5.7% of the UK population [84]. THIN was shown to be broadly representative of the UK population in terms of demographics, prevalence of major conditions, and death rates adjusted for demographics and deprivation [86].

Figure 6.1 outlines the structure of the data files available in THIN. Data are ordered according to the practice level, followed by the patient level. For each practice, there are seven main files including patient, medical, therapy, additional health data (AHD), postcode variable indicators (PVI), consultation, and staff files; all of which are linked by unique practice and patient identifiers. In addition, medical, therapy and AHD files are linked to consultation and staff files by consultation and staff identifiers.

Data in THIN are mainly recorded as coded information. Clinical data (e.g. symptoms and medical diagnoses) are recorded using Read codes, a hierarchical coding system used to document clinical summary information in primary care [87, 88]. Each entry in the medical record can be accompanied by free text comments, as the Vision software allows for the entry of free text or scanned information. Prescriptions are entered using multilex drug codes, which are categorised as per chapters in the British National Formulary [89]. Additional health information (e.g. lifestyle factors and health indicator measurements) is coded using the additional health data (AHD) codes. Information on deprivation is given by quintiles of the Townsend score [90], which is an indicator of deprivation in the patients' postcode and is calculated based on unemployment, house and car ownership, and household overcrowding. In THIN, Townsend deprivation score is coded as a five-level categorical variable, with level 1 corresponding to quintile 1 (least deprived), and level 5 corresponding to quintile 5 (most deprived).

In terms of quality assurance, the acceptable mortality reporting (AMR) date [91] was calculated and applied to each practice in THIN; this information is supplied for every practice. The AMR date is the date after which the practice is deemed to be reporting a rate of all-cause mortality sufficiently similar to that expected for a practice with the same demographics, based on data from the Office for National Statistics (ONS) [91]. The AMR date can also be combined with other measures of data reporting quality such as the acceptable computer usage (ACU) date

Figure 6.1. Structure of the main data files for each participating general practice in The Health Improvement Network (THIN) database.



[92]. The ACU date is designed to exclude the transition period between the practice switching from paper-based records to complete computerisation. It is defined as the date from which the practice is consistently recording on average at least two drug prescriptions, one medical record and one additional health record per patient per year [92].

6.2.2 *The use of primary care databases in research*

Data in primary care databases are a valuable resource, and are increasingly tapped into for use in epidemiological and health research for several reasons. First, the data are collected in an unobtrusive and automatic way, and therefore closely represent the real-life primary care data recording practice. Second, the use of medical data collected by general practices across the UK provides not only representation of the local population, but the wide geographical coverage of the databases also allows for broader generalisation to the overall UK population [86, 93]. Third, as patient data for some practices date back to the early 1990s, the databases provide access to a rich source of longitudinal health data for research, making it possible to conduct analysis

over the patients' lifetime. In addition, the large size of the databases offers a relatively cheaper, faster, and more accessible alternative for research on populations that are otherwise logistically difficult and expensive to enrol in clinical trials or standard observational studies. This include, for example, research on low incidence/prevalence diseases, individuals with severe mental illness, [1, 2], pregnant women [3–5], children [6], and the very elderly [7, 8]. The databases have also increasingly been linked to secondary care and mortality datasets.

Despite their unique potentials for use in research, primary care databases are not free from limitations. For studies based on these data sources, it is important to have good measures of key variables such as height, body weight, blood pressure, cholesterol level, smoking and alcohol consumption, many of which are routinely recorded in primary care. However, since information is mainly collected in primary care for the purposes of clinical management, non-trivial amounts of data among these variables are often missing, posing challenges to analysis and inference [94–97].

6.2.3 *Data recording in primary care and the issue of missing data*

Data in primary care are collected mainly as a means for healthcare staff to track information about their patients' health, such as required to diagnose or monitor a condition. Therefore, physicians might not find it necessary, or they might not have enough time during the consultations to record all information that is relevant for research. As a result, the data recording practice in primary care is directly reflected in how the data are present in primary care databases. A missing value in a patient's record might indicate that either the patient did not attend the GP consultation, or the value was not measured during the consultation, or the value was measured but for some reason (e.g. incorrect measurement) was not recorded. The second scenario implies that the missing data may not technically be considered as missing, because they are usually not intended to be recorded. In order to apply existing methods for handling missing data, it is assumed that such values would have been observed for all individuals if requested by the GPs.

There have been some changes in the data recording in primary care over the last two decades. Efforts were made by the NHS to improve the recording of several health indicators such as height, body weight, blood pressure, smoking, and alcohol consumption in primary care through the implementation of some pay-for-performance initiatives. The New Patient Health Checks scheme [98], which was first introduced under the NHS Department of Health contract between GPs and the government back in 1990, provided GPs with incentive payments to collect health indicator measurements for all newly registered individuals. This scheme eventually came to an end in 2004 when the Quality and Outcomes Framework (QOF) was introduced under the revised Department of Health contract [99]. Under QOF, practices receive financial rewards for high quality recording of health indicators that are required to monitor specific clinical conditions, e.g. smoking status recorded in the preceding 15 months for diabetic patients [100]. Data for health indicators associated with QOF-specific diseases have therefore been recorded much more regularly [96, 97].

The implementation of these financial incentives, together with the fact that information in primary care is mainly recorded for clinical purposes, suggest that missing data in primary care databases are not likely to be missing completely at random (MCAR). Indeed, Delaney et al.

[101] examined data in the GPRD database and reported that individuals with more frequent blood pressure readings tended to have higher recorded values. Similarly, Marston et al. [95] explored the recording of several health indicators for newly registered individuals in THIN between 2004 and 2006 and compared that to external nationally representative datasets. Their findings were consistent with data at practice registration being missing at random (MAR) for height, body weight, and blood pressure, whereas missing not at random (MNAR) for smoking status and alcohol consumption. The 2008 update of QOF on the recording of smoking status was likely to start improving the completeness of this variable. Using THIN data, Marston et al. [96] later re-examined the recording of smoking status for newly registered individuals between 2008 and 2009 and concluded that the variable was relatively well recorded for these individuals. Following from their findings, Marston et al. [96] reported that if information on smoking status was missing for an individual, then he or she was likely to be an ex-smoker or a non-smoker [96].

6.3 ETHNICITY RECORDING IN PRIMARY CARE

Disparities in health and healthcare among racial and ethnic groups are an issue of growing interest. Recognition of this comes from a body of research, which increasingly uses electronic health records from clinical and administrative health databases [12, 102]. Ethnicity recording has been incorporated in UK primary care, and therefore ethnicity information is also available in a number of large primary care databases [12]. However, as is often the case with using such databases where data are mainly collected for patient care management, research addressing ethnicity is hindered by the low level of recording [13, 15, 94, 103]. In a recent study investigating the recording of ethnicity in the CPRD database, Mathur et al. [13] showed that less than 30% of individuals had their ethnicity recorded between 1990 and 2012. Although the completeness of ethnicity data in primary care has improved for newly registered patients after the financial incentivisation to record ethnicity was introduced under QOF between 2006/7 and 2011/12 [13], some practices may still not record ethnicity on a regular basis.

Despite incomplete ethnicity information being a pervasive problem in research using electronic health records, it is unclear how to handle missing data in ethnicity when ethnicity is either an outcome of interest or a risk factor for health outcomes. In practice, some previous studies omitted ethnicity from the main analysis due to a high level of missing values [2]. Excluding ethnicity from the analysis will obscure its effect and is also likely to confound the associations of other variables in the analysis. Another widely used method for missing data in ethnicity, which is also the default option in most statistical software, is complete record analysis (CRA, section 2.3.1), where the analysis is performed only on individuals with fully observed data on all variables included in the analysis. Consider the setting where missing data only occur in ethnicity. It is known that a CRA gives valid inferences when ethnicity is MCAR. In analyses examining the association between ethnicity and an outcome of interest, CRA can also provide unbiased estimates in settings where missing data in ethnicity are not MCAR, provided that missingness in ethnicity is conditionally independent of the outcome, given ethnicity and/or other fully observed variables [39, 40]. However, even when this assumption holds, analysis based only on the complete records can still be inefficient, since information available in individuals with missing ethnicity is discarded. In order to retain all individuals in the analysis, some previous studies

using UK primary care databases made an assumption that individuals with missing values in ethnicity belonged to the White ethnic group, and replaced missing values with the White ethnicity [19]. The assumption underpinning this approach is that only individuals from the White ethnic group ever failed to have their ethnicity recorded, which might not be plausible in practice. One consequence of this single imputation approach is that some non-White individuals can be incorrectly assigned to the White ethnic group, which can bias any effect of the non-White ethnic groups on the outcome. In addition, variances are also likely to be underestimated (section 2.3.2).

Large primary care databases are evolving in recent years, and are increasingly enriched by record linkage to different data sources. In particular, databases such as the CPRD or THIN are now linked to the Hospital Episode Statistics. It is a secondary care data warehouse containing information recorded during a patient's time in the hospital; it also includes ethnicity information. This linkage therefore allows for some additional ethnicity information recorded in secondary care to be recovered in the primary care records. The possibility for data linkage depends on whether the practice is willing to participate in the linkage scheme, and in THIN database this is only available for a few practices in England [104, 105]. Data linkage algorithms based on patient identifiers (such as the NHS numbers) can also be prone to linkage errors [106]. In addition to record linkage, previous work exploring the use of name-recognition software to indirectly retrieve ethnicity information of the patients showed that the credibility of this method is questionable, particularly for the non-White ethnic groups and descendants of migrants [13, 107, 108].

Multiple imputation (MI, section 2.4) is increasingly applied for accommodating missing data in studies using primary care databases in recent years [2, 19, 95, 96, 109]. MI offers a more statistically sensible approach for dealing with missing data compared to other simple 'ad-hoc' methods, and MI requires researchers to think carefully about the plausible assumptions for the missingness mechanism. Despite gaining more popularity in practice, the main obstacle for MI in large clinical databases is the limited available information on the extent and mechanisms that give rise to missing data, since procedures and incentives for recording data in primary care change over time [95]. As a result, the missingness mechanism assumptions made by standard MI might not be plausible, which can in turn affect the method's validity. In particular, although the standard implementation of MI in most statistical packages assumes data are MAR, at any particular time in a dynamic database the MAR assumption might not be plausible. The probability that ethnicity is recorded in primary care may well vary systematically with ethnicity, even after its associations with other variables are taken into account. The recording of ethnicity information can also be related to other factors which are not available in routine healthcare databases such as patients' circumstances during consultation or at admission, the availability of staff, the lack of time or opportunity to ask the patients about their ethnicity [13]. This implies a potential MNAR mechanism for ethnicity, and as a result, standard MI might not be an appropriate approach to missing data. Standard MI might fail to yield plausible estimation of the marginal distribution of ethnicity, and can potentially distort the association between ethnicity and the health outcome in the main analysis.

Although MI can be extended to impute missing values under the MNAR mechanism (section

2.5), imputation becomes more difficult because a model for the missing data mechanism needs to be specified, which describes how missingness depends on both observed and unobserved quantities. This implies that in practice, it is necessary to define a model for either the probability of observing a variable and its unseen values (the selection model [24]), or the difference in the distribution of individuals with and without missing data (the pattern-mixture model [64, 65]). The extra model specification requirement in MI under the MNAR mechanism raises several issues. First, the underlying MAR and MNAR mechanisms are not verifiable from the observed data alone. Second, there can be an infinite number of possible MNAR models for any dataset, and it is very rare to know which of these models is appropriate for the missingness mechanism. Due to the potential complexity of modelling the MNAR mechanism, analyses assuming data are MNAR are relatively infrequently performed and reported in the applied literature. In practice, researchers more often try to enhance the plausibility of the MAR assumption as much as possible by the inclusion of many (auxiliary) variables in the imputation model [47, 110].

For some variables in certain datasets, their corresponding population marginal distributions can be obtained from external data sources, such as population censuses or surveys. If our study samples come from such a population, it is natural to feed the population information into the imputation process in order to calibrate inference to the population. For ethnicity information recorded in UK primary care databases, the distribution of ethnicity in the UK population is available in the UK population census. Since the majority of the UK population is registered with a GP, the population-level distribution of ethnicity can be utilised in MI of missing data in ethnicity in primary care databases to inform the imputation. Previous chapters of this thesis propose and evaluate two candidate population-calibrated MI methods which exploit such external information: *weighted MI* and *calibrated-δ adjustment MI*. In these approaches, the population distribution of the incomplete variable can be used to calculate appropriate probability weights or a δ adjustment in the imputation model's intercept, which are then used in MI such that the post-imputation distribution much more closely (and often exactly) matches the population level.

The rest of this chapter presents two case studies in which the issue of incomplete ethnicity information in UK primary care databases is used to demonstrate the application of weighted MI and calibrated-δ adjustment MI, as well as to compare weighted MI and calibrated-δ adjustment MI to existing methods for handling missing ethnicity data in practice. The first case study focuses on ethnicity, where it is of interest to estimate the marginal distribution of ethnicity in UK primary care databases. In the second case study, the focus is on examining the association between ethnicity and the prevalence of type 2 diabetes diagnoses in UK primary care, where ethnicity is included as a covariate in the analysis model.

6.4 Case study 1: assessing the missing at random assumption for ethnicity in the health improvement network primary care database

The first case study is a cross-sectional study which aims to assess the plausibility of the MAR assumption for ethnicity in UK primary care databases. Since the population marginal distribution of ethnicity is available in the UK census data, the MAR assumption for ethnicity can be assessed by using standard MI to handle missing data and comparing the resulting ethnicity distribution

to that in the census. Mathur et al. [13] previously compared the 2011 UK census distribution of ethnicity to that among individuals with a record of ethnicity who were actively registered in the CPRD database on the census day. It was concluded that there was not much discrepancy between the two sources.

The objective of this study is to compare the distribution of ethnicity in THIN general practices in London with that in the 2011 UK census data for London, after missing values in ethnicity are handled by (i) a CRA, (ii) single imputation with the White ethnic group, (iii) standard MI assuming MAR, (iv) marginal and (v) conditional weighted MI, and (vi) calibrated-δ adjustment MI. In the population-calibrated MI methods, the 2011 UK census distribution of ethnicity in London [111] is used as the reference distribution.

As discussed in section 6.2.3, similar comparisons between data recorded in THIN and external nationally representative datasets were explored by Marston et al. [95, 96] for health indicators including height, weight, smoking status, and alcohol consumption. As an example of this, Marston et al. [96] reported that if smoking status is missing for an individual then (s)he is typically either an ex-smoker or non-smoker, and accordingly proposed only allowing imputed data to take one of these two values. As illustrated in chapters 3 and 4, the population-calibrated MI methods supersede this approach, providing a way to incorporate population distribution information into MI.

6.4.1 Study sample

Data used in this study are from individuals who are permanently registered (i.e. variable `patflag` takes value A or C) with general practices contributing data to THIN which meet the data quality assurance criteria (AMR [91] and ACU [92], section 6.2.1). From this population, a sample of all individuals registered with general practices in London is selected for subsequent analyses. This sample is chosen since it is not only more practical to perform MI on a smaller dataset, but also because London is the most ethnically diverse region in the UK, and hence incorrect assignment of ethnicity from imputing missing data with the White ethnic group is expected to be more apparent compared to other regions.

All individuals actively registered with THIN general practices in London on the 2011 UK census day (27 March 2011) are identified. For each individual, a start date is defined as the latest of [date of birth, ACU and AMR dates, registration date], and an end date is defined as the earliest of [date of death, date of transfer out of practice, and date of last data collection from the practice]. Individuals are selected into the study sample if their start date is on or before the 2011 UK census day, and their end date is on or after the census day.

6.4.2 Outcome variable

Ethnicity is typically recorded in THIN using the Read code system [87]; it can also be recorded using free text entries. A Read code list including codes related to ethnicity is developed using a published method [88] (appendix D.1). The majority of ethnicity records are identified by searching both the medical and additional health data files for Read codes in the ethnicity code list. Additional information is gathered by searching both the pre-anonymised free text, as well as other free text linked to ethnicity-related Read codes. Ethnicity is then coded into the five-level

ONS classification as White, Mixed, Asian, Black, and Other ethnic groups [112]. Subsequently, the Mixed and Other ethnic groups are combined due to the small numbers of individuals in these two groups. Searching for ethnicity-related Read codes reveals that there are individuals with multiple records of ethnicity, some of which are inconsistent. For these individuals, it can not be determined with certainty whether their ethnicity is in fact one of the recorded categories or if all the recorded categories are incorrect. Therefore, their ethnicity is set to missing for simplicity, since the issue of inconsistency in ethnicity recording is not the focus of this study.

6.4.3 *Statistical analysis*

For MI of ethnicity, a multinomial logistic regression imputation model is constructed for ethnicity using information on individuals' age in 2011, sex, and quintiles of the Townsend deprivation score. Age is analysed in 10-year age groups for individuals aged 0-79 years, and all individuals aged 80 years and older are grouped into the 80+ category. In addition, indicators of common diseases known to be associated with ethnicity including heart attack, stroke, type 2 diabetes, chronic kidney disease, sickle cell disease, thalassemia, and schizophrenia are also included in the imputation model. These variables are chosen after consultations with two GP colleagues (Dr Claudia Cooper and Dr Kate Walters). This is done because the patterns of ethnicity might differ by demographic variables and between those with and without each disease [47]. In order to extract disease information in THIN, a Read code list is developed for each disease, and both medical and additional health data files are searched for Read codes in the corresponding disease code list. Since the aim of this study is to illustrate the use of marginal and conditional weighted MI and calibrated-δ adjustment MI in a univariate missing data setting where only ethnicity contains missing values, individuals with missing data in other variables (year of birth, sex, Townsend score) are excluded from the analysis.

In this study, ethnicity is analysed as a four-level categorical variable (White, Asian, Black, Mixed/Other). Therefore, the univariate calibrated-δ adjustment MI method for handling missing data in an incomplete binary variable discussed in section 4.2 can be generalised for handling missing values in ethnicity as a categorical variable. The overall proportion of the j^{th} level of ethnicity, $j = 1, \ldots, 4$, can be written as

$$p\left(\text{ethnicity} = j\right) = p\left(\text{ethnicity} = j \mid r = 0\right) p\left(r = 0\right) + p\left(\text{ethnicity} = j \mid r = 1\right) p\left(r = 1\right), \quad (6.1)$$

where $p\left(\text{ethnicity} = j\right)$ is obtained from the census data; $p\left(r = 0\right)$, $p\left(\text{ethnicity} = j \mid r = 1\right)$, and $p\left(r = 1\right)$ are available in the observed data.

A multinomial logistic regression imputation model for ethnicity conditional on 10-year age groups (0–9 years old as the base level), sex (male as the base level), Townsend deprivation score (quintile 1 as the base level), binary indicators of heart attack, stroke, type 2 diabetes, chronic kidney disease, sickle cell disease, thalassemia, and schizophrenia (no diagnosis as the base level) is fitted to the observed data. Setting the first level of ethnicity (White, $j = 1$) as the base level to identify the model, the probability of the level j^{th} of ethnicity in the observed data, $j = 2, \ldots, 4$, can be written in terms of the observed-data linear predictors, $\text{linpred}_j^{\text{obs}}$, obtained from the

multinomial logistic regression model for ethnicity as

$$p\left(\text{ethnicity} = j \mid r = 1\right) = \frac{1}{n^{\text{obs}}} \sum_{i=1}^{n^{\text{obs}}} \frac{1}{1 + \sum_{j=2}^{4} \exp\left(\text{linpred}_{ij}^{\text{obs}}\right)}, \tag{6.2}$$

where i indexes individuals in the dataset, and

$$\begin{aligned}
\text{linpred}_{ij}^{\text{obs}} = \theta_{j\text{o}}^{\text{obs}} &+ \sum_{a=10}^{80} \theta_{j\text{age}_a}^{\text{obs}} I\left[\text{age}_{ij} = a\right] + \theta_{j\text{fem}}^{\text{obs}} I\left[\text{sex}_{ij} = \text{female}\right] \\
&+ \sum_{t=2}^{5} \theta_{j\text{town}_t}^{\text{obs}} I\left[\text{Townsend}_{ij} = t\right] + \theta_{j\text{ha}}^{\text{obs}} I\left[\text{heart attack}_{ij} = \text{yes}\right] \\
&+ \theta_{j\text{str}}^{\text{obs}} I\left[\text{stroke}_{ij} = \text{yes}\right] + \theta_{j\text{t2d}}^{\text{obs}} I\left[\text{type 2 diabetes}_{ij} = \text{yes}\right] \\
&+ \theta_{j\text{ckd}}^{\text{obs}} I\left[\text{kidney disease}_{ij} = \text{yes}\right] + \theta_{j\text{sic}}^{\text{obs}} I\left[\text{sickle cell}_{ij} = \text{yes}\right] \\
&+ \theta_{j\text{tha}}^{\text{obs}} I\left[\text{thalassemia}_{ij} = \text{yes}\right] + \theta_{j\text{sch}}^{\text{obs}} I\left[\text{schizophrenia}_{ij} = \text{yes}\right],
\end{aligned} \tag{6.3}$$

where $I\left[\,\right]$ denotes the indicator function taking values 1 if the statement inside the brackets is true and 0 otherwise.

Following the methods outlined in section 4.2, since covariates in the imputation model for ethnicity are all binary or categorical, the log odds ratios are the same among those with ethnicity observed and missing. The linear predictors in the missing data, $\text{linpred}_{j}^{\text{mis}}$, can therefore be written as

$$\begin{aligned}
\text{linpred}_{ij}^{\text{mis}} = \left(\theta_{j\text{o}}^{\text{obs}} + \delta_{j\text{o}}\right) &+ \sum_{a=10}^{80} \theta_{j\text{age}_a}^{\text{obs}} I\left[\text{age}_{ij} = a\right] + \theta_{j\text{fem}}^{\text{obs}} I\left[\text{sex}_{ij} = \text{female}\right] \\
&+ \sum_{t=2}^{5} \theta_{j\text{town}_t}^{\text{obs}} I\left[\text{Townsend}_{ij} = t\right] + \theta_{j\text{ha}}^{\text{obs}} I\left[\text{heart attack}_{ij} = \text{yes}\right] \\
&+ \theta_{j\text{str}}^{\text{obs}} I\left[\text{stroke}_{ij} = \text{yes}\right] + \theta_{j\text{t2d}}^{\text{obs}} I\left[\text{type 2 diabetes}_{ij} = \text{yes}\right] \\
&+ \theta_{j\text{ckd}}^{\text{obs}} I\left[\text{kidney disease}_{ij} = \text{yes}\right] + \theta_{j\text{sic}}^{\text{obs}} I\left[\text{sickle cell}_{ij} = \text{yes}\right] \\
&+ \theta_{j\text{tha}}^{\text{obs}} I\left[\text{thalassemia}_{ij} = \text{yes}\right] + \theta_{j\text{sch}}^{\text{obs}} I\left[\text{schizophrenia}_{ij} = \text{yes}\right],
\end{aligned} \tag{6.4}$$

where $\delta_{j\text{o}}$ is the level–j intercept adjustment in the multinomial logistic regression imputation model for ethnicity. Hence, the probability of the j^{th} level of ethnicity in the missing data, $j = 2, \ldots, 4$, is given by

$$p\left(\text{ethnicity} = j \mid r = 0\right) = \frac{1}{n^{\text{mis}}} \sum_{i=1}^{n^{\text{mis}}} \frac{1}{1 + \sum_{j=2}^{4} \exp\left(\text{linpred}_{ij}^{\text{mis}}\right)}. \tag{6.5}$$

From (6.1)–(6.5), the problem now becomes finding the solutions $\delta_{j\text{o}}, j = 2, \ldots, 4$, of a system of three non-linear equations for the three categories of ethnicity. Instead of using interval bisection which is no longer sufficient in this case because there is a system of equations to be solved simultaneously, the solutions can be obtained by using the Stata command `nl` [113] and defining a function evaluator program.

All MI methods are performed using $M = 30$ imputations ($\approx$ percentage missing ethnicity), and Rubin's rules [20, 21] are used to obtain overall estimates of the ethnic proportions and associated standard errors. All analyses are performed using Stata 14 [44], where `mi impute mlogit` is used for standard MI, my command `mi impute wmlogit` for marginal and conditional weighted MI, `mi impute mlogit [pweight]` for calibrated-δ adjustment MI, and `mi`

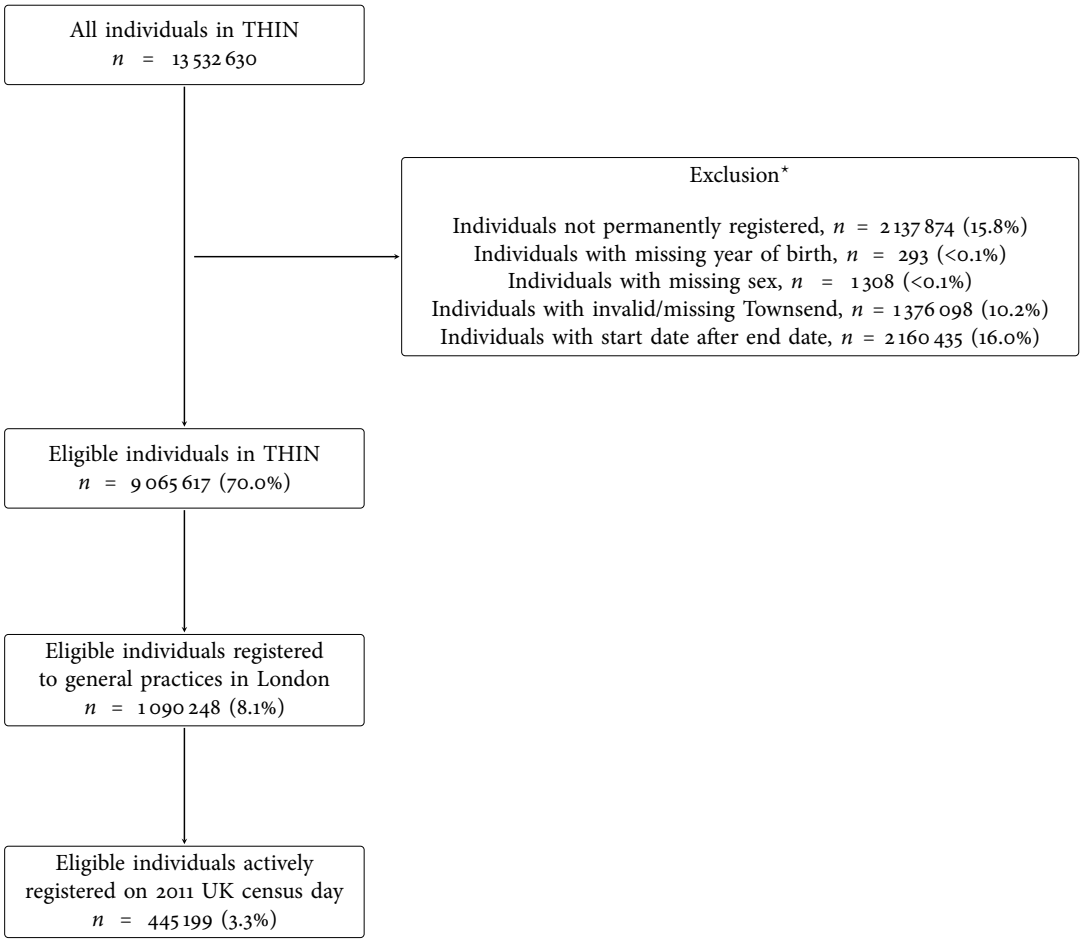
`estimate: proportion` for performing the main analysis in completed datasets and obtaining the final results using Rubin's rules [20, 21].

6.4.4 *Results*

Figure 6.2 depicts a flowchart of the selection criteria used to select the relevant sample for this study. Data from a total of $n = 13\,532\,630$ individuals are extracted from THIN, of which $n = 2\,137\,874$ (15.8%) individuals are not permanently registered (variable `patflag` does not take values A or C), $n = 293$ (< 0.1%) individuals do not have their year of birth recorded, $n = 1\,308$ (< 0.1%) individuals have missing sex, $n = 1\,376\,098$ (10.2%) individuals have an invalid or missing Townsend deprivation score, and $n = 2\,160\,435$ (16.0%) individuals have their start date after their end date. Applying the selection criteria to the data results in $n = 9\,065\,617$ (70.0%) individuals eligible for inclusion in this study. Among these eligible individuals, there are $n = 1\,090\,248$ (8.1%) individuals who are registered to THIN general practices in London, of whom $n = 445\,199$ (3.3%) individuals are actively registered on the 2011 UK census day and make up the THIN sample relevant to this study.

Table 6.1 presents a summary of variables used in this study. The sample comprises 51% female; the majority of individuals in the sample (more than 80%) are below 60 years of age;

Figure 6.2. Case study 1: flowchart of selection criteria for THIN sample.



* Note: an individual can be excluded from the study sample due to more than one criterion.

slightly more than 70% of individuals have a Townsend score quintile of 3 and above; type 2 diabetes and chronic kidney disease are the two most frequently diagnosed conditions, followed by stroke and heart attack, schizophrenia, thalassemia, and sickle cell disease.

Ethnicity is recorded for 337 278 (75.8%) and missing for 107 921 (24.2%) individuals (table 6.2). Among individuals with ethnicity recorded, the percentage of the White ethnic group is higher and the percentages of the non-White ethnic groups are lower compared to the ethnic breakdown in the 2011 UK census (table 6.2). Single imputation of missing values in ethnicity with the White ethnic group exaggerates this discrepancy, further overestimating the White ethnic group and underestimating other non-White groups (table 6.2).

Table 6.1. Case study 1: summary of variables in the analysis; n = 445 199.

Variable	Frequency	%
Sex		
Male	219 071	49.21
Female	226 128	50.79
Age group (years)		
0–9	51 472	11.56
10–19	47 444	10.66
20–29	62 047	13.94
30–39	76 855	17.26
40–49	70 576	15.85
50–59	53 270	11.97
60–69	39 642	8.90
70–79	25 401	5.71
80+	18 492	4.15
Townsend score		
Quintile 1 (least deprived)	51 770	11.63
Quintile 2	68 643	15.42
Quintile 3	109 545	24.61
Quintile 4	113 417	25.48
Quintile 5 (most deprived)	101 824	22.87
Disease indicator		
Heart attack	5 865	1.32
Stroke	8 990	2.02
Type 2 diabetes	26 235	5.89
Chronic kidney disease	21 000	4.72
Sickle cell disease	331	0.07
Thalassemia	2 458	0.55
Schizophrenia	2 360	0.53

Table 6.2. Case study 1: distribution of ethnicity when missing values are included, excluded, and singly imputed with the White ethnic group; n = 445 199.

Ethnicity	Frequency	% including missing	% excluding missing	Frequency missing imputed as White	% missing imputed as White	% 2011 UK census London
White	245 064	55.05	72.66	352 985	79.29	59.80
Asian	37 519	8.43	11.12	37 519	8.43	18.50
Black	33 374	7.50	9.90	33 374	7.50	13.30
Mixed/Other	21 321	4.79	6.32	21 321	4.79	8.40
Missing	107 921	24.24				
$\sum$ including missing	445 199					
$\sum$ excluding missing	337 278					

Table 6.3 presents the adjusted associations of ethnicity (among individuals with ethnicity recorded) with fully observed variables included as covariates in the imputation model for ethnicity, including sex, age group, Townsend deprivation score, and disease indicators. Corresponding unadjusted associations are shown in table D.2 (appendix D.2). Adjusted relative risk ratios (RRR) and 95% confidence intervals (CIs, table 6.3) are estimated by fitting a multivariable multinomial logistic regression of four-level ethnicity conditional on fully observed variables among individuals with observed ethnicity; the White ethnic group is set as the base level for ethnicity.

Table 6.4 presents the adjusted associations of missingness in ethnicity with fully observed variables included as covariates in the imputation model for ethnicity. Corresponding unadjusted associations are given in table D.3 (appendix D.2). A multivariable logistic regression model for the response indicator of ethnicity conditional on fully observed variables is fitted to the whole study sample to obtain adjusted odds ratios (OR) and 95% CIs (table 6.4).

These results suggest that sex, age group, Townsend score, and disease indicators are related to ethnicity. Apart from sickle cell disease and thalassemia, these variables are also associated with missingness in ethnicity, supporting the inclusion of these variables as covariates in the imputation model for ethnicity [46, 114].

Table 6.3. Case study 1: adjusted associations of ethnicity with variables used to inform the imputation of ethnicity among the complete records; n = 337 278.

	Asian		Black		Mixed/Other	
Variable	RRR	95% CI	RRR	95% CI	RRR	95% CI
Sex						
Male	1		1		1	
Female	0.905	0.885; 0.926	1.106	1.080; 1.132	1.036	1.007; 1.066
Age group (years)						
0–9	1		1		1	
10–19	0.929	0.887; 0.974	1.047	1.001; 1.096	0.916	0.869; 0.965
20–29	0.895	0.860; 0.932	0.502	0.480; 0.524	0.605	0.576; 0.634
30–39	0.885	0.852; 0.920	0.531	0.509; 0.553	0.552	0.527; 0.578
40–49	0.602	0.577; 0.627	0.728	0.699; 0.758	0.492	0.468; 0.517
50–59	0.498	0.475; 0.522	0.504	0.481; 0.529	0.402	0.379; 0.426
60–69	0.350	0.332; 0.370	0.225	0.210; 0.240	0.227	0.211; 0.245
70–79	0.318	0.298; 0.340	0.286	0.266; 0.307	0.172	0.155; 0.190
80+	0.136	0.123; 0.150	0.094	0.084; 0.106	0.078	0.066; 0.092
Townsend score						
Quintile 1 (least deprived)	1		1		1	
Quintile 2	1.261	1.196; 1.330	1.263	1.174; 1.360	1.221	1.135; 1.314
Quintile 3	2.277	2.173; 2.386	2.756	2.585; 2.937	2.117	1.984; 2.258
Quintile 4	2.480	2.367; 2.598	4.739	4.455; 5.041	2.719	2.552; 2.897
Quintile 5 (most deprived)	2.457	2.343; 2.577	8.049	7.571; 8.558	3.948	3.706; 4.204
Disease indicator						
Heart attack	1.187	1.073; 1.314	0.466	0.393; 0.554	0.875	0.726; 1.055
Stroke	0.851	0.772; 0.939	0.935	0.838; 1.044	0.784	0.666; 0.924
Type 2 diabetes	3.455	3.310; 3.606	2.140	2.033; 2.252	1.555	1.445; 1.672
Chronic kidney diseases	1.052	0.988; 1.120	1.325	1.239; 1.416	1.014	0.919; 1.119
Sickle cell disease	1.991	0.631; 6.282	128.666	69.419; 238.480	12.231	5.515; 27.124
Thalassemia	6.666	5.963; 7.453	3.569	3.096; 4.114	4.754	4.088; 5.530
Schizophrenia	0.830	0.706; 0.977	1.916	1.698; 2.162	1.031	0.843; 1.259

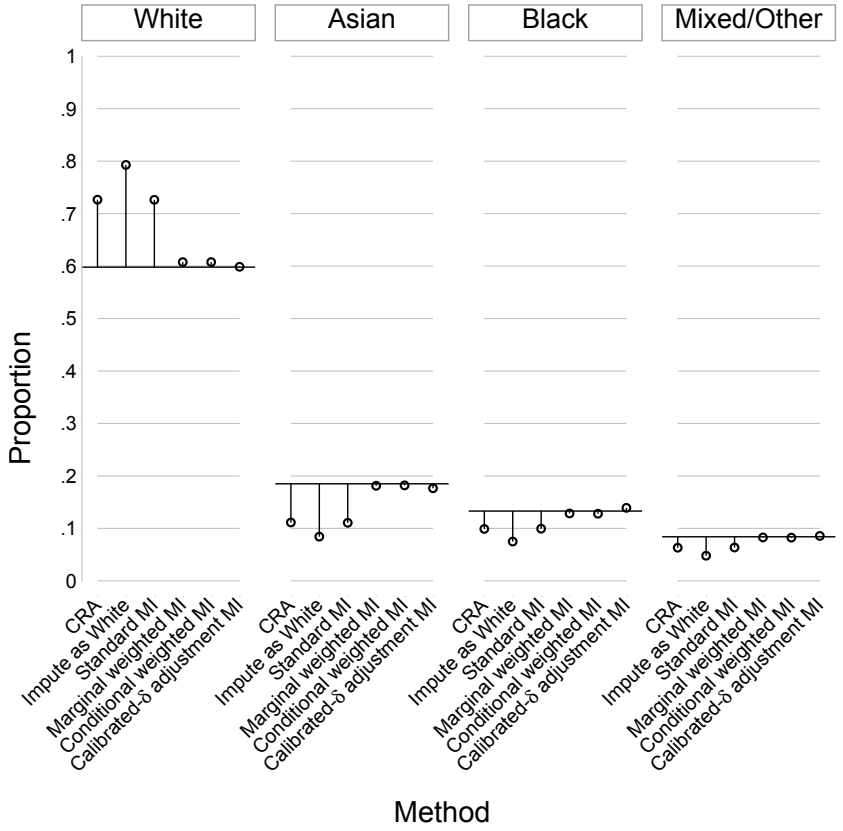
* Note: White ethnic group and no diagnosis are base levels for ethnicity and disease indicators, respectively; RRR: relative risk ratios are obtained from fitting a multivariable multinomial logistic regression model for four-level ethnicity conditional on all variables considered; CI: confidence interval.

Table 6.4. Case study 1: adjusted associations of the response indicator of ethnicity with variables used to inform the imputation of ethnicity, n = 445 199.

	OR	95% CI
Sex		
Male	1	
Female	1.310	1.291; 1.328
Age group (years)		
0–9	1	
10–19	0.512	0.498; 0.527
20–29	0.975	0.948; 1.003
30–39	1.162	1.131; 1.195
40–49	0.828	0.806; 0.850
50–59	0.744	0.723; 0.766
60–69	0.863	0.836; 0.891
70–79	0.923	0.888; 0.959
80+	0.851	0.815; 0.890
Townsend score		
Quintile 1 (least deprived)	1	
Quintile 2	0.969	0.943; 0.995
Quintile 3	0.908	0.886; 0.930
Quintile 4	1.070	1.044; 1.097
Quintile 5 (most deprived)	0.962	0.938; 0.987
Disease indicator		
Heart attack	1.283	1.199; 1.373
Stroke	1.140	1.079; 1.204
Type 2 diabetes	1.400	1.354; 1.448
Kidney disease	1.179	1.134; 1.225
Sickle cell disease	0.978	0.759; 1.261
Thalassemia	1.008	0.917; 1.109
Schizophrenia	1.347	1.213; 1.496

* Note: no diagnosis is the base level for disease indicators; OR: odds ratios are obtained from fitting a multivariable logistic regression model for the response indicator of ethnicity conditional on all variables considered; CI: confidence interval.

Figure 6.3. Case study 1: distribution of four-level ethnicity in different methods for handling missing ethnicity data.



* Note: horizontal black lines: the 2011 UK census ethnic breakdown for London is used as the reference distribution for marginal and conditional weighted MI and calibrated-δ adjustment MI, as well as for comparison.

Figure 6.3 presents the distribution of four-level ethnicity obtained from the various methods for missing data in ethnicity. Standard errors in all methods are very small (less than 0.0009), hence 95% CIs are not shown. As seen in table 6.2, compared to the 2011 UK census statistics for London, the White ethnic group is over-represented among the complete records. It is apparent that single imputation of missing ethnicity values with the White ethnic group further overestimates the proportion of the White group, while underestimating the Asian, Black, and Mixed/Other groups, under the assumption that the ethnicity distribution in THIN should match the census. Standard MI produces similar proportion estimates to that in CRA (figure 6.3). Despite the inclusions of several variables which are thought to be predictive of both the values of ethnicity and missingness in ethnicity in the imputation model for ethnicity, the distribution of ethnicity after standard MI still does not match that in the census. This result suggests that ethnicity might be MNAR even after conditioning on variables in the imputation model. Marginal and conditional weighted MI and calibrated-δ adjustment MI compensate for the over-representation of the White ethnic group in the observed data by imputing missing values with this group less frequently, in order to give the correct census proportions in the completed data. Conversely, the non-White ethnic groups, which are under-represented in the observed data, are imputed more frequently. As a results, the weighted MI and calibrated-δ adjustment MI methods yield ethnic proportion estimates that are closest to the census level.

Table 6.5. Case study 1: fraction of missing information (Monte Carlo error) for the estimated proportions of ethnicity.

Ethnicity	Standard MI	Marginal weighted MI	Conditional weighted MI	Calibrated-δ adjustment MI
White	0.147 (0.027)	0.174 (0.039)	0.172 (0.038)	0.217 (0.046)
Asian	0.175 (0.043)	0.451 (0.075)	0.445 (0.076)	0.378 (0.066)
Black	0.159 (0.036)	0.352 (0.070)	0.356 (0.073)	0.478 (0.055)
Mixed/Other	0.247 (0.051)	0.401 (0.073)	0.401 (0.071)	0.411 (0.052)

Table 6.5 presents the fraction of missing information (FMI) in the various methods for missing data in ethnicity. Higher FMI is obtained in marginal and conditional weighted MI and calibrated-δ adjustment MI compared to standard MI. This could be explained by the fact that non-White ethnic groups, which are under-represented in the observed data, are imputed more often in marginal and conditional weighted MI and calibrated-δ adjustment MI than in standard MI. Therefore, the between-imputation variance relies on more imputed values in the non-White ethnic groups and less frequently imputed values in the White ethnic group, which leads to the non-White proportion estimates being more variable across the completed datasets. This might be the reason why the FMI is higher in the population-calibrated MI methods in comparison to standard MI.

6.5 Case study 2: ethnicity and the prevalence of type 2 diabetes diagnoses in the health improvement network primary care database

While the first case study focuses on summarising the overall distribution of ethnicity as the outcome, in practice it is often of interest to consider ethnicity as a covariate in the main analysis. For this reason, this second case study aims to illustrate the use of the marginal and conditional weighted MI and calibrated-δ adjustment MI methods for handling missing data in ethnicity, when ethnicity is included as a covariate in the analysis model. In particular, case study 2 is a cross-sectional study which examines the association between ethnicity and the prevalence of type 2 diabetes diagnoses in THIN database in 2013. Prevalence of type 2 diabetes is chosen as the outcome variable to illustrate the application of the population-calibrated MI methods as developed in chapters 3 and 4.

The objective of this study is to examine how the prevalence of type 2 diabetes diagnoses in THIN varies with ethnicity, after adjusting for individuals' demographics including age in 2013, sex, and social deprivation measured by quintiles of the Townsend score.

6.5.1 Study sample

All individuals who are permanently registered (variable `patflag` takes values A or C) with general practices in London contributing data to THIN are considered for inclusion in the study sample, for the reasons explained in section 6.4.1. For each individual, a start date is defined as the latest of (date of birth, ACU and AMR dates, registration date). Similarly, an end date is defined as the earliest of (date of death, date of transfer out of practice, date of last data collection

from the practice). Point prevalence of type 2 diabetes on 01 January 2013 is calculated, since THIN is a dynamic database in which individuals register with and leave their general practices at different times. Individuals are selected into the study sample if they are actively registered to THIN practices on 01 January 2013, and in addition they need to have been registered with the same general practices for at least 12 months by 01 January 2013. This criterion is introduced to ensure that there is enough time for the individuals to have their type 2 diabetes recorded in their electronic health data, from when the individuals first register with their general practices.

6.5.2 *Outcome variable*

The recording of diabetes diagnoses and management in THIN is comprehensive and therefore there are several ways an individual may be identified as diabetic. For this study, an algorithm developed by Sharma et al. [115] is used to identify individuals with diabetes mellitus, as well as to distinguish between type 1 and type 2 diabetes. According to this algorithm, individuals are identified as having diabetes if they have at least two of the following records: a diagnostic code for diabetes, supporting evidence of diabetes (e.g. screening for diabetic retinophany), or prescribed treatment for diabetes [115, 116]. In this study, the first record of any of these three is considered as the date of diagnosis. In addition to identifying individuals with diabetes, the algorithm also distinguishes between type 1 and type 2 diabetes based on individuals' age at diagnosis, types of treatment and timing of the diabetes diagnosis [115, 116]. See Sharma et al. [115] for more information about the algorithm. After the study sample is selected using the method described in section 6.5.1, prevalent cases of type 2 diabetes are defined according to the point prevalence approach, in which all individuals who have a diagnosis of type 2 diabetes by 01 January 2013 are defined as prevalent cases.

6.5.3 *Statistical analysis*

The analysis model in this study is a logistic regression model for a binary indicator of whether an individual has a diagnosis of type 2 diabetes on or before 01 January 2013, conditional on the individual's ethnicity (defined as four categories), age in 2013, sex, and Townsend deprivation score (defined in quintiles). Age is analysed in 10-year age groups for individuals aged 0-79 years, and all individuals aged 80 years and older are grouped into the 80+ category. Ethnicity information is extracted and categorised as described in section 6.4.2.

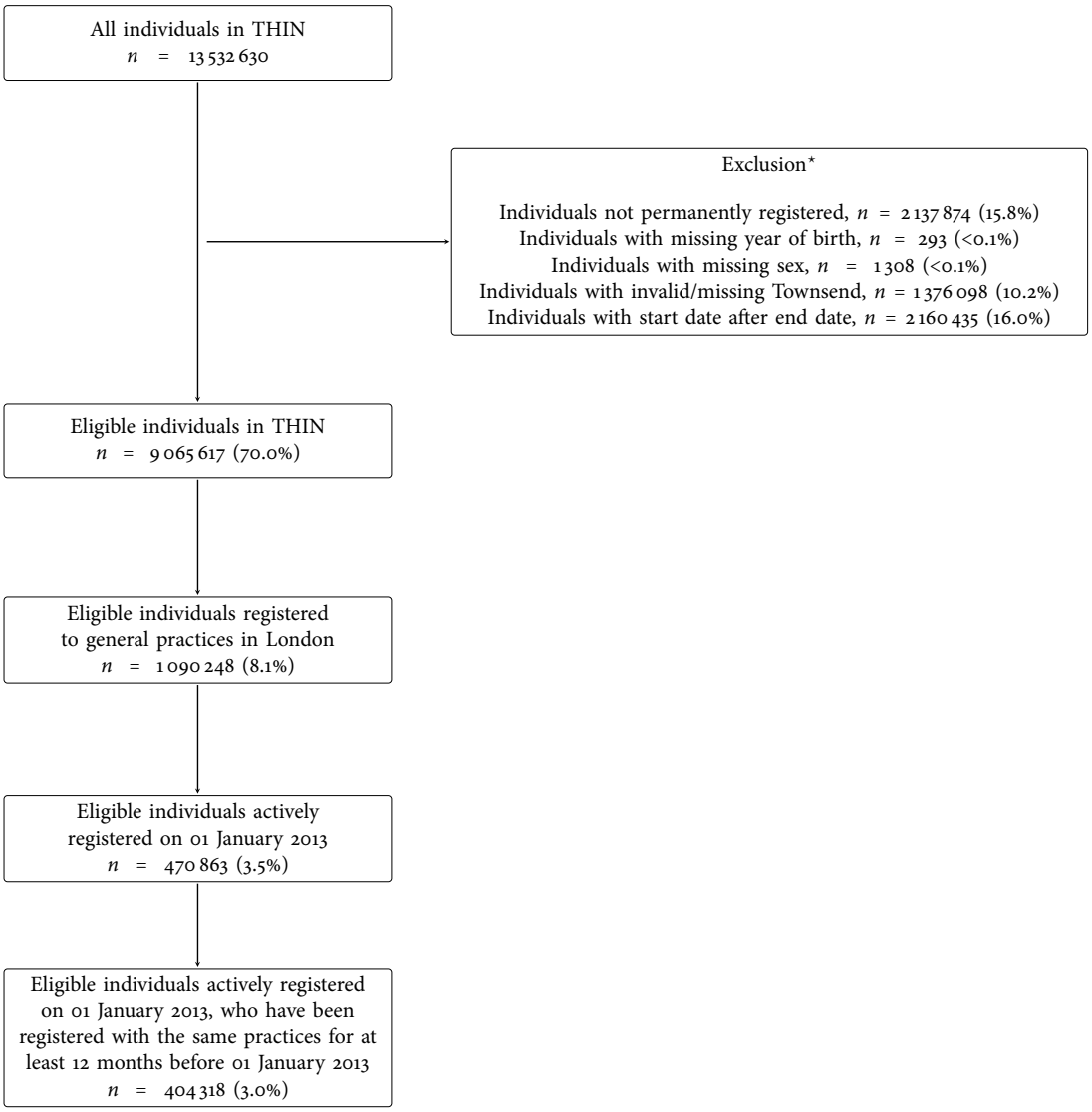
Missing values in ethnicity are handled by (i) a CRA, (ii) single imputation with the White ethnic group, (iii) standard MI, (iv) marginal and (v) conditional weighted MI, and (vi) calibrated-δ adjustment MI using the 2011 UK census distribution of ethnicity in London [111] as the reference distribution. For MI of ethnicity, a multinomial logistic regression imputation model is constructed for ethnicity using all variables in the analysis model, including individuals' age group in 2013, sex, and quintiles of the Townsend deprivation score. In MI, the outcome variable must be explicitly included in the imputation model for the incomplete covariate [47]. Since the analysis model is a logistic regression model, the binary indicator of whether an individual has a diagnosis of type 2 diabetes in 2013 is also included as a covariate in the imputation model for ethnicity. Calibrated-δ adjustment MI is performed using the same procedure as outlined in section 6.4.3, with the relevant covariates in the imputation model for ethnicity in this study.

All MI methods are performed using $M = 30$ imputations, and Rubin's rules [20, 21] are used to obtain estimates of associations and standard errors. All analyses are conducted using Stata 14 [44], where `mi impute mlogit` is used for standard MI, my command `mi impute wmlogit` [73] for marginal and conditional weighted MI, `mi impute mlogit [pweight]` for calibrated-δ adjustment MI, and `mi estimate: logit` for performing the main analysis in completed datasets and obtaining the final results using Rubin's rules [20, 21].

6.5.4 *Results*

Figure 6.4 depicts a flowchart of the selection criteria used to select the relevant sample for this study. In total, data from $n = 13\,532\,630$ individuals are extracted from THIN, of which $n = 2\,137\,874$ (15.8%) individuals are not permanently registered (variable `patflag` does not take values A or C), $n = 293$ (< 0.1%) individuals do not have their year of birth recorded, $n = 1\,308$ (< 0.1%) individuals have missing sex, $n = 1\,376\,098$ (10.2%) individuals have invalid or

Figure 6.4. Case study 2: flowchart of selection criteria for THIN sample.



* Note: an individual can be excluded from the study sample due to more than one criterion.

missing Townsend deprivation score, and $n = 2\,160\,435$ (16.0%) individuals have their start date after their end date. Applying the selection criteria results in $n = 9\,065\,617$ (70.0%) individuals eligible for inclusion in this study. Among the eligible individuals, there are $n = 1\,090\,248$ (8.1%) individuals who are registered to THIN general practices in London, of whom $n = 470\,863$ (3.5%) individuals are actively registered on 01 January 2013. Finally, $n = 404\,318$ (3.0%) individuals have at least 12 months of follow-up by 01 January 2013 and make up the sample for this study.

Table 6.6 presents a summary of variables used in this study. The sample comprises 51% female; the majority of individuals in the sample (approximately 80%) are below 60 years of age; slightly more than 70% of the individuals have a Townsend score quintile of 3 and above; and 5.5% of the individuals have a diagnosis of type 2 diabetes on or before 01 January 2013.

Ethnicity is recorded for 309 684 (76.6%) and missing for 94 634 (23.4%) individuals (table 6.7). Among individuals with ethnicity recorded, the estimated proportion of the White ethnic group is higher, and the non-White ethnic groups lower compared to the corresponding ethnic breakdown in the 2011 UK census data for London (table 6.7). Single imputation with the White ethnic group further overestimates the White group and underestimates the other non-White groups, under the assumption that the ethnicity distribution in THIN should match the census (table 6.7).

Table 6.8 presents the adjusted associations of ethnicity (among individuals with ethnicity recorded) and fully observed variables in the analysis model, including sex, age group, Townsend deprivation score, and an indicator of type 2 diabetes. Corresponding unadjusted associations are shown in table D.4 (appendix D.2). Relative risk ratios (RRR) and 95% CIs in table 6.8 are obtained from fitting a multivariable multinomial logistic regression model for four-level ethnicity conditional on fully observed variables, with the White ethnic group set as the base level for ethnicity.

Table 6.9 presents the adjusted associations of missingness in ethnicity and fully observed variables in the analysis model. Corresponding unadjusted associations are given in table D.5 (appendix D.2). Odds ratios (OR) and 95% CIs in table 6.9 are obtained from fitting a multivariable logistic regression model for the response indicator of ethnicity, conditional on fully observed variables.

These results suggest that sex, age group, Townsend score, and the indicator of type 2 diabetes are related to both the chance of observing ethnicity as well as the ethnic groups, supporting the inclusion of these variables as covariates in the imputation model for ethnicity [46, 114].

Table 6.6. Case study 2: summary of variables in the analysis; n = 404 318.

Variable	Frequency	%
Sex		
Male	198 301	49.05
Female	206 017	50.95
Age group (years)		
0–9	41 601	10.29
10–19	45 664	11.29
20–29	50 065	12.38
30–39	65 695	16.25
40–49	64 837	16.04
50–59	53 272	13.18
60–69	39 427	9.75
70–79	25 348	6.27
80+	18 409	4.55
Townsend score		
Quintile 1 (most deprived)	48 934	12.10
Quintile 2	64 788	16.02
Quintile 3	101 305	25.06
Quintile 4	102 626	25.38
Quintile 5 (least deprived)	86 665	21.43
Type 2 diabetes	22 100	5.47

Table 6.7. Case study 2: distribution of ethnicity when missing values are included, excluded, and singly imputed with the White ethnic group; n = 404 318.

Ethnicity	Frequency	% including missing	% excluding missing	Frequency missing imputed with White	% missing imputed with White	% 2011 UK census London
White	224 403	55.5	72.46	319 037	78.91	59.8
Asian	35 027	8.66	11.31	35 027	8.66	18.8
Black	30 771	7.61	9.94	30 771	7.61	13.3
Other	19 483	4.82	6.29	19 483	4.82	8.4
Missing	94 634	23.41				
$\sum$ including missing	404 318					
$\sum$ excluding missing	309 684					

Table 6.8. Case study 2: adjusted associations of ethnicity with variables used to inform the imputation of ethnicity among the complete records; n = 309 684.

	Asian		Black		Mixed/Other	
Variable	RRR	95% CI	RRR	95% CI	RRR	95% CI
Sex						
Male	1		1		1	
Female	0.925	0.904; 0.947	1.131	1.103; 1.159	1.061	1.030; 1.094
Age group (years)						
0–9	1.480	1.416; 1.546	1.329	1.272; 1.388	1.969	1.869; 2.075
10–19	1.463	1.397; 1.532	1.520	1.454; 1.589	1.894	1.793; 2.001
20–29	1.363	1.307; 1.422	0.812	0.777; 0.850	1.227	1.162; 1.296
30–39	1.424	1.371; 1.480	0.711	0.682; 0.741	1.089	1.035; 1.146
40–49	1		1		1	
50–59	0.797	0.762; 0.834	0.835	0.799; 0.873	0.809	0.762; 0.858
60–69	0.580	0.551; 0.611	0.340	0.319; 0.361	0.483	0.449; 0.520
70–79	0.516	0.485; 0.548	0.421	0.394; 0.450	0.369	0.335; 0.407
80+	0.262	0.240; 0.285	0.180	0.162; 0.199	0.168	0.145; 0.195
Townsend score						
Quintile 1 (least deprived)	1		1		1	
Quintile 2	1.251	1.186; 1.320	1.305	1.211; 1.406	1.249	1.159; 1.346
Quintile 3	2.308	2.202; 2.419	2.825	2.647; 3.016	2.129	1.993; 2.275
Quintile 4	2.481	2.368; 2.600	4.836	4.540; 5.152	2.708	2.538; 2.889
Quintile 5 (most deprived)	2.442	2.327; 2.564	8.708	8.179; 9.272	4.081	3.826; 4.354
Type 2 diabetes	3.561	3.405; 3.724	2.244	2.126; 2.368	1.620	1.499; 1.751

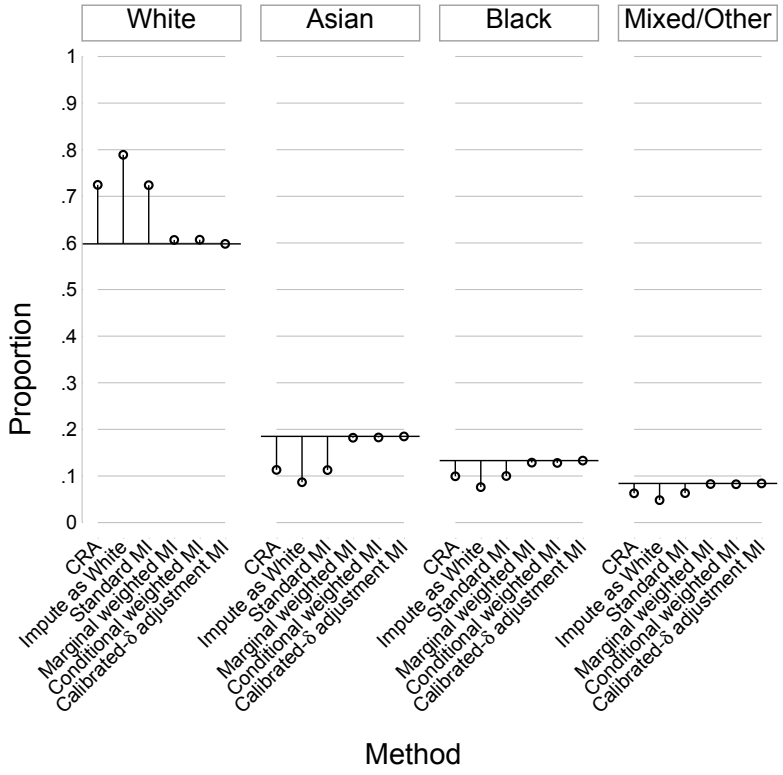
* Note: White ethnic group and no diagnosis are base levels for ethnicity and type 2 diabetes, respectively; RRR: relative risk ratios are obtained from fitting a multivariable multinomial logistic regression model for four-level ethnicity conditional on all variables considered; CI: confidence interval.

Table 6.9. Case study 2: adjusted associations of the response indicator of ethnicity with variables used to inform the imputation of ethnicity; n = 404 318.

Variable	OR	95% CI
Sex		
Male	1	
Female	1.258	1.240; 1.277
Age group (years)		
0–9	1.263	1.225; 1.302
10–19	0.549	0.535; 0.564
20–29	0.895	0.871; 0.920
30–39	1.343	1.307; 1.380
40–49	1	
50–59	0.845	0.823; 0.868
60–69	0.945	0.917; 0.974
70–79	1.090	1.051; 1.130
80+	1.050	1.008; 1.093
Townsend score		
Quintile 1 (least deprived)	1	
Quintile 2	0.964	0.937; 0.991
Quintile 3	0.906	0.884; 0.930
Quintile 4	1.072	1.045; 1.101
Quintile 5 (most deprived)	0.941	0.916; 0.966
Type 2 diabetes	1.389	1.340; 1.441

* Note: no diagnosis is the base level for type 2 diabetes; OR: odds ratios are obtained from fitting a multivariable logistic regression model for the response indicator of ethnicity conditional on all variables considered; CI: confidence interval.

Figure 6.5. Case study 2: distribution of four-level ethnicity in different methods for handling missing ethnicity data.

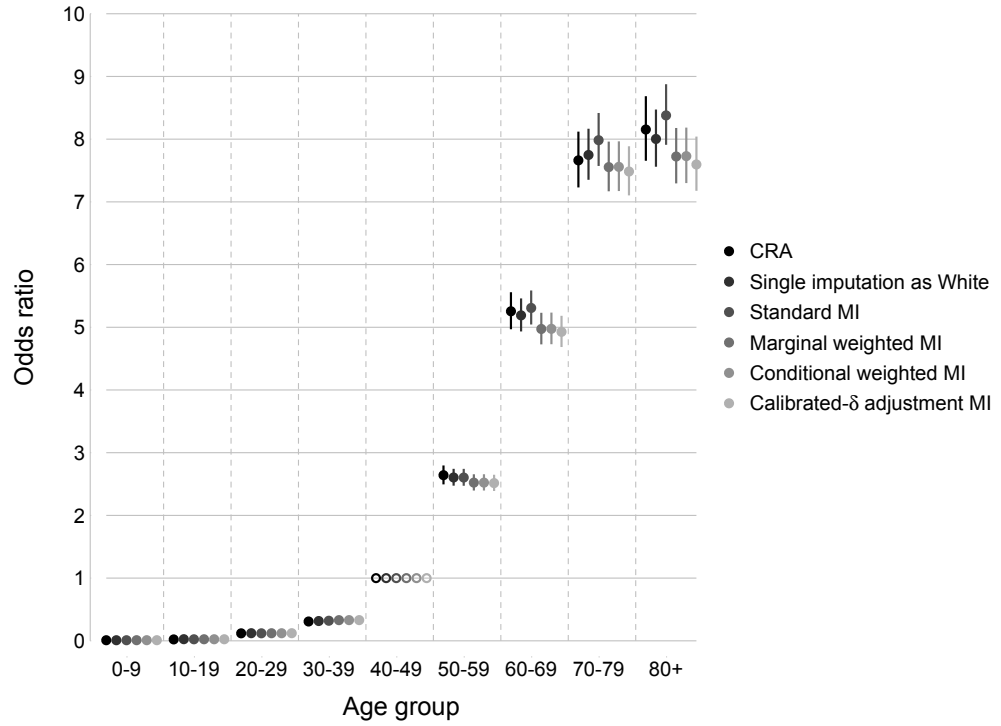


* Note: horizontal black lines: the 2011 UK census ethnic breakdown for London is used as the reference distribution for marginal and conditional weighted MI and calibrated-δ adjustment MI, as well as for comparison.

Figure 6.5 shows the distribution of four-level ethnicity after missing values in ethnicity are handled by the various methods for missing data. As before, CRA, single imputation of missing values with the White ethnic group, and standard MI overestimate the White group while underestimating the other non-White ethnic proportions, compared to the corresponding 2011 UK census statistics. In marginal and conditional weighted MI and calibrated-δ adjustment MI, the majority of missing values in ethnicity are imputed with the Asian and Black groups. These methods recover the ethnic breakdown in the census as expected, since the census distribution is used as the reference for these population-calibrated MI methods.

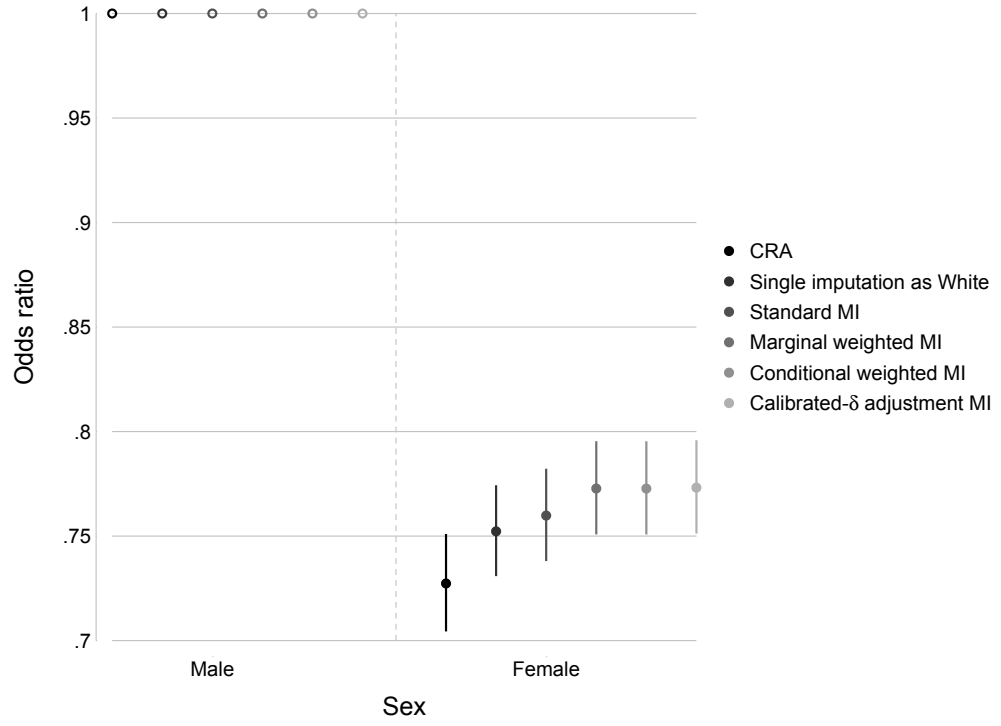
Figures 6.6–6.9 present estimated odds ratios of type 2 diabetes prevalence and 95% CIs for age group, sex, Townsend deprivation score, and ethnicity in the analysis model. Age 40–49 years, male, quintile 1, and the White ethnic group are selected as base levels for age group, sex, Townsend score, and ethnicity, respectively. $M = 30$ imputations produce Monte Carlo errors for point estimates of less than 10% of the estimated standard errors. The relative efficiency versus an infinite number of imputations is > 0.988 for all parameter estimates and MI methods. Overall, the odds of being diagnosed with type 2 diabetes increases relatively smoothly with older age groups and higher quintiles of the Townsend deprivation score; is lower in female compared to male; and is higher in the Asian, Black, and Mixed/Other ethnic groups compared to the White group in all methods for handling missing data in ethnicity.

Figure 6.6. Case study 2: estimated odds ratio of type 2 diabetes diagnosis for age group in different methods for handling missing ethnicity data.



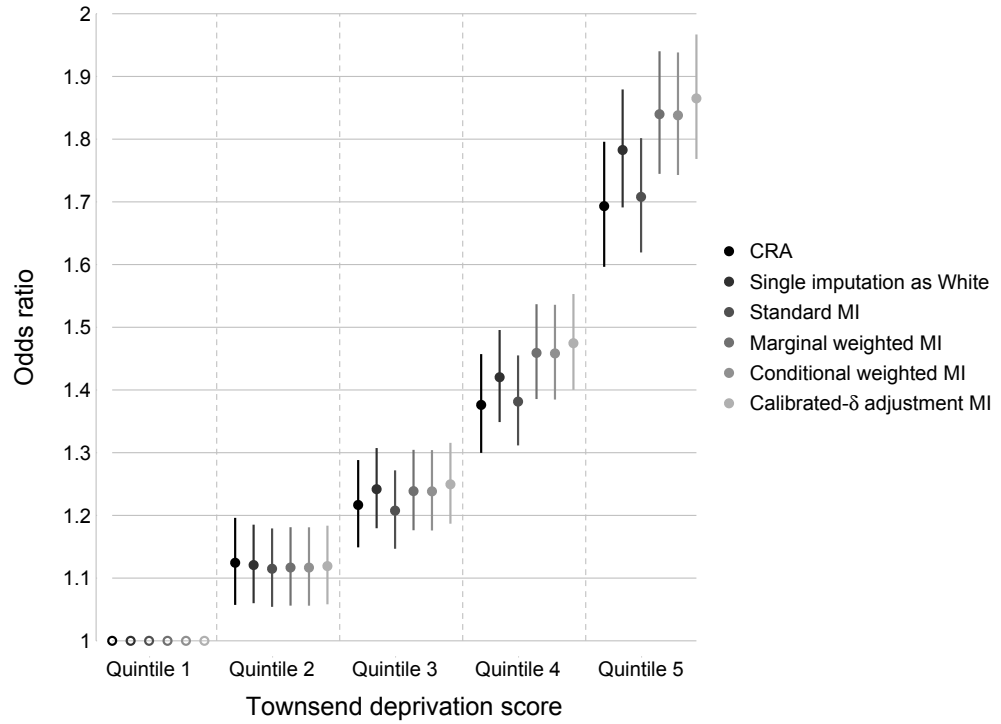
* Note: hollow circles: 40–49 years is the base level for age group; error bars: 95% confidence intervals.

Figure 6.7. Case study 2: estimated odds ratio of type 2 diabetes diagnosis for sex in different methods for handling missing ethnicity data.



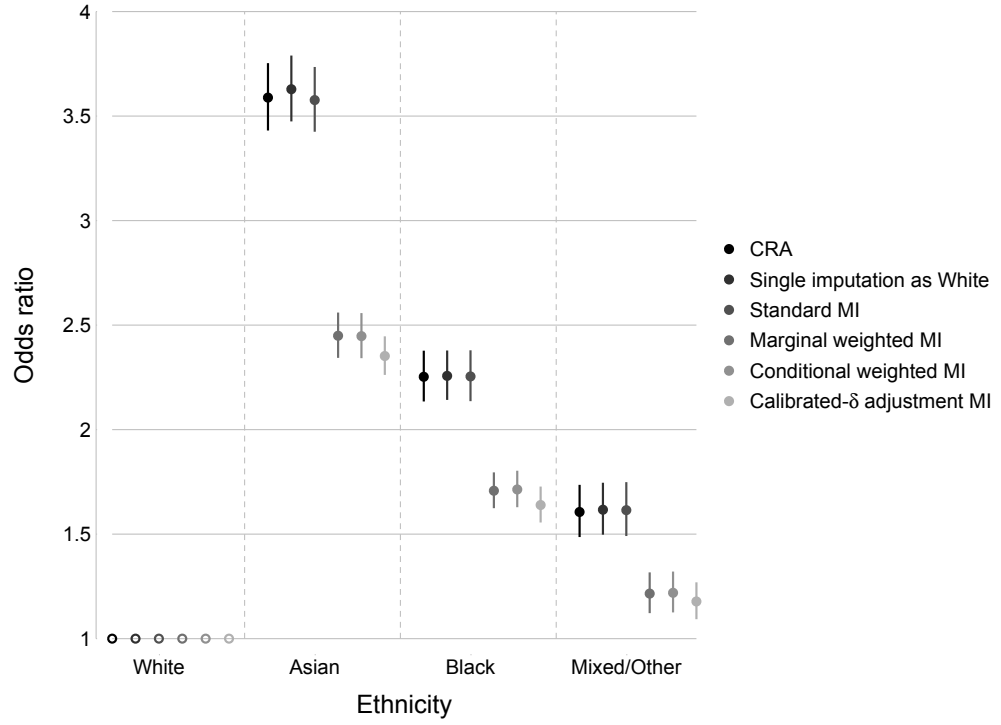
* Note: hollow circles: male is the base level for sex; error bars: 95% confidence intervals.

Figure 6.8. Case study 2: estimated odds ratio of type 2 diabetes diagnosis for Townsend deprivation score in different methods for handling missing ethnicity data.



* Note: hollow circles: quintile 1 (least deprived) is the base level for Townsend score; error bars: 95% confidence intervals.

Figure 6.9. Case study 2: estimated odds ratio of type 2 diabetes diagnosis for ethnic group in different methods for handling missing ethnicity data.



* Note: hollow circles: White ethnic group is the base level for ethnicity; error bars: 95% confidence intervals.

Marginal and conditional weighted MI produce broadly similar results. Compared to CRA, single imputation with the White ethnic group, and standard MI, the population-calibrated MI methods (marginal and conditional weighted MI and calibrated-δ adjustment MI) produce comparable estimated odds ratios for the younger age groups, and smaller estimated odds ratios for the older (60+) age groups. The population-calibrated MI methods lead to slightly higher estimated odds ratio for female compared to CRA, single imputation with the White ethnic group, and standard MI, and this increase is towards the null. All missing data methods yield odds ratios that increase with more deprived quintiles of the Townsend score. Again, the population-calibrated MI methods yield similar estimated odds ratios compared to the other methods for the three lower quintiles of the Townsend score, and higher estimates for higher deprivation scores.

The most noticeable differences in point estimates associated with the prevalence of type 2 diabetes diagnoses are seen in the estimated odds ratios for ethnicity. CRA, single imputation, and standard MI again return similar results, in which the odds of having a diagnosis of type 2 diabetes is around 3.6 times higher in the Asian ethnic group compared to the White group, and individuals in the Black ethnic group are about 2.25 times more likely to receive a diagnosis of type 2 diabetes compared to those of White ethnic background. Singly imputing missing ethnicity values with the White ethnic group slightly increases the estimated odds ratios for non-White ethnic groups. This is because individuals with missing ethnicity are, on average, less likely to have a diagnosis of type 2 diabetes (table 6.9). Replacing missing values with the White ethnic group means that this group will contain a lower percentage of type 2 diabetes diagnoses, which implies that the estimated odds ratios for the non-White ethnic groups will increase. Compared to CRA, single imputation with the White ethnic group, and standard MI, the population-calibrated MI methods lead to a reduction in estimated odds ratios for the non-White ethnic groups, with point estimates being slightly lower in calibrated-δ adjustment MI compared to the two weighted MI approaches. For all non-White ethnic groups, the 95% CIs of point estimates in the population-calibrated MI methods do not cross that of the other methods.

In the population-calibrated MI methods where missing values are more frequently imputed with the non-White ethnic groups, the explanatory power of ethnicity for type 2 diabetes is diluted, with lower estimates of odds ratios for the non-White ethnic groups. These findings might correspond to the stronger effect of the Townsend score, which compensates for the reduction in the odds ratios for ethnicity. The odds ratios for Townsend score are smaller in CRA compared to population-calibrated MI, for higher deprivation quintiles. In addition, these findings seem to suggest that some effect of ethnicity is absorbed in Townsend score in the population-calibrated MI methods, where Townsend score explains some of the effect which might otherwise be explained by ethnicity. This might be attributed to a possibility that individuals of Asian and Black ethnic groups, whose ethnicity is not recorded, tend to belong to the more deprived quintiles of the Townsend score.

Returning to the missingness mechanisms considered thus far for the development of marginal and conditional weighted MI and calibrated-δ adjustment MI, results in case study 2 suggest a potential departure from the MAR assumption for missingness in ethnicity. This is because conditional on the outcome variable type 2 diabetes and other fully observed variables

Table 6.10. Case study 2: fraction of missing information (Monte Carlo error) for the estimates of association between ethnicity and the prevalence of type 2 diabetes diagnoses.

Method	Asian	Black	Mixed/Other
Standard MI	0.132 (0.033)	0.193 (0.050)	0.230 (0.066)
Marginal weighted MI	0.355 (0.066)	0.173 (0.039)	0.342 (0.060)
Conditional weighted MI	0.351 (0.065)	0.195 (0.042)	0.338 (0.058)
Calibrated-δ adjustment MI	0.186 (0.027)	0.259 (0.062)	0.251 (0.074)

* Note: White ethnic group is the base level for ethnicity.

included in the analysis model, standard MI does not yield a distribution of ethnicity that is comparable to the census ethnic breakdown. Ethnicity is also not likely to be MNAR dependent on only the values of ethnicity, since the point estimates in CRA and standard MI are broadly comparable.

Results from analyses exploring the associations between covariates in the imputation model for ethnicity and missingness in ethnicity suggest that sex, age, Townsend deprivation score, and type 2 diabetes are factors likely to be related to whether ethnicity is recorded. This indicates that ethnicity might be MNAR depending on the ethnic groups, fully observed outcome (type 2 diabetes), as well as other fully observed covariates in the analysis model (sex, age group, Townsend score).

Table 6.10 presents the fraction of missing information (FMI) for the estimates of association between ethnicity and the prevalence of type 2 diabetes diagnoses in different MI methods for handling missing data in ethnicity. Again, standard MI tends to have lower FMI compared to the population-calibrated MI methods, which is consistent with the results seen in case study 1.

6.6 SUMMARY

This chapter provides a description of The Health Improvement Network (THIN) database which is the main data source for the case studies presented in this chapter. Section 6.2 explains how data are typically recorded in the primary care setting, which gives rise to the issue of missing data in primary care databases. Since the development of the population-calibrated MI methods is motivated by the issue of missing data in ethnicity in research using primary care databases, the recording of ethnicity information in primary care is discussed in section 6.3. These are followed by two case studies in sections 6.4 and 6.5, which illustrate the application of marginal and conditional weighted MI and calibrated-δ adjustment MI in utilising population-level external information about the marginal distribution of ethnicity for handling missing values in ethnicity. These methods are also compared to standard MI, single imputation of missing values with the White ethnic group, and CRA in the case studies.

The first case study (section 6.4) aims to assess the plausibility of the MAR assumption for ethnicity, by estimating the distribution of ethnicity and comparing that to the 2011 UK census statistics. It is shown that CRA, single imputation of missing values with the White ethnic group, and standard MI can lead to implausible distributions of ethnicity, whereas the population-calibrated MI methods can yield ethnicity distributions that are comparable to that given by the census statistics. The second case study (section 6.5) examines the association between ethnicity

and the prevalence of type 2 diabetes diagnoses, where ethnicity is considered as a covariate in the analysis model with some missing values. Compared to CRA, single imputation of missing values with the White ethnic group, and standard MI, both marginal and conditional weighted MI and calibrated-δ adjustment MI methods result in scientifically relevant changes in inferences, typically for the Asian and Black ethnic groups. Potential missingness mechanisms for missing values in ethnicity are discussed and also compared to the missingness mechanisms considered in univariate simulation studies presented in chapters 3 and 4.

The next chapter provides an overall discussion of the methodology and applications developed and presented thus far to conclude this thesis.

7

Discussion

7.1 SUMMARY OF THESIS

It is very difficult to completely avoid the problem of missing data in research involving human participants, and therefore missing data commonly occur in medical research and epidemiological studies. The presence of missing data is even harder to control in research using electronic health record databases of patients' clinical information, since such databases are not designed for this purpose. Missing data can hinder researchers from performing standard statistical analyses

designed for complete datasets, and can potentially lead to bias and inefficiency.

There is a vast range of methods for accommodating missing data, most of which assume data are missing at random (MAR, section 2.2). The MAR assumption simplifies analyses with missing data, such that analyses can proceed without having to explicitly define a model for the missingness mechanism. However, in many real-life settings it is possible that data are missing not at random (MNAR, section 2.2). In addition, since the validity of the MAR assumption cannot be verified, it is important to consider other plausible assumptions for the missingness mechanism underlying the unseen values, which reflect potential departure from MAR towards MNAR. An analysis based on the MAR assumption should therefore be accompanied by a sensitivity analysis exploring how inferences may vary under alternative assumptions about the potential MNAR mechanism [117]. Unfortunately, sensitivity analyses are often not performed or reported sufficiently in practice [30, 118], a tendency abetted by the practical constraints of many applied projects.

This thesis explored the idea of calibrating the dataset used for analysis to a relevant population-level external data source, as a means of anchoring the analysis to the population in the presence of missing data suspected to be MNAR. The investigations conducted in this thesis centred around the use of multiple imputation (MI) [20] (section 2.4) for handling missing values in incomplete covariates in the analysis model of interest. In practice, MI is widely implemented under the assumption of data being MAR. Although the standard implementation of MI provides a good starting point for the analysis when data are suspected to be MNAR, the assumption made about the missingness mechanism is not correct. One indication of potential departure from the MAR assumption towards a MNAR mechanism is when standard MI results in an implausible marginal distribution of the incomplete variable.

MI offers flexibility for performing sensitivity analyses, since the imputation model can be tuned to incorporate possible departure from the MAR assumption [47, 117]. However, such tuning is typically 'un-anchored', as it is hard to decide what the sensitivity parameters should be or to justify the choice of values for the sensitivity parameters. This motivated the idea of bringing in information from external data sources into the imputation process, in order to improve standard MI when there are reasons to believe that data are MNAR. More specifically, for an incomplete variable in the analysis dataset, if its corresponding population marginal distribution is available in an external dataset, knowledge about such a distribution can be integrated into MI in order to match the incomplete variable's post-imputation distribution to that in the population. The rationale for aligning the post-imputation distribution with the population reference is that the imputation should be performed consistently with the population data.

Two population-calibrated univariate MI approaches for utilising the incomplete variable's population marginal distribution in the imputation process were proposed, evaluated, and compared to existing methods for handling missing data. The first approach is *(marginal and conditional) weighted multiple imputation*. This approach involves weighting the complete records in the imputation model with probability weights which are calculated based on the incomplete variable's population distribution. The second approach is *calibrated-δ adjustment multiple imputation*. In this approach, the intercept of the imputation model is adjusted by an offset which is derived using the incomplete variable's population distribution as the reference.

The development of the two population-calibrated univariate MI methods focused on missing values in an incomplete binary/categorical covariate in the analysis model. Several analytic and simulation studies were conducted to evaluate these methods under a range of missingness mechanisms with an increasing level of complexity (chapters 3–5). These univariate MI methods were also incorporated in the multivariate imputation by chained equations (MICE) algorithm [23] for imputing missing values in more than one incomplete covariate, accounting for their population marginal distributions which are available externally (sections 3.6 and 4.4). The application of these methods was illustrated in case studies using real-life data from a large UK primary care electronic health record database (sections 6.4 and 6.5).

Findings from previous chapters of this thesis are summarised in the following sections.

7.1.1 Weighted multiple imputation of a binary covariate when the outcome variable is binary

Chapter 3 proposed and evaluated the univariate *weighted multiple imputation* method for an incomplete binary/categorical covariate. Weighted MI augments the standard MI method (under the MAR assumption) with sampling probability weights. These weights are derived using population-level information about the incomplete covariate's distribution from an external dataset, in order to match the incomplete covariate's post-MI distribution to that in the population.

An outline of the weighted MI procedure for an incomplete binary/categorical variable, followed by the derivation of the marginal and conditional weights in weighted MI were provided. Univariate analytic and simulation studies were conducted to evaluate and compare marginal and conditional weighted MI to standard MI and complete record analysis (CRA) in terms of bias and other finite-sample properties in a univariate missing data setting where missing values occur in a single covariate (sections 3.3 and 3.4). These studies featured a 2×2 contingency table with a fully observed binary outcome variable y and a partially observed covariate x, where the analysis model was a logistic regression of y on x. Overall, empirical bias agreed closely with what the theoretical calculations predicted. The four missingness mechanisms considered are listed below, together with the methods which produced unbiased estimates of both parameters of the analysis model under each scenario.

1. x was MCAR: all methods were valid;
2. x was MAR conditional on y: standard MI and conditional weighted MI were valid;
3. x was MNAR dependent on x: CRA and marginal weighted MI were valid;
4. x was MNAR dependent on x and y: none of the methods were valid; conditional weighted MI appeared to be the least biased method.

Results over repeated simulations showed that when bias was present, coverage of 95% confidence intervals (CI) was lower than the nominal level and efficiency was lower than the full data (i.e. before values in x are set to missing). When a method being evaluated was unbiased under a posited missingness mechanism, correct coverage and matching standard errors were often achieved. The key finding to be taken forward to the development of calibrated-δ adjustment MI (chapter 4) was that the effects of covariates in the imputation model for the incomplete covariate need to be accounted for in the derivation of the weights, and conditional weights might not account for such effects in an optimal manner.

Chapter 3 also investigated the setting where the population distribution of the incomplete

variable used to derive the weights in weighted MI is not 'known', i.e. it is not obtained from a population census or equivalent (section 3.5). When there is uncertainty in estimating the population distribution, a natural approach to incorporate this extra source of uncertainty in the imputation process would be to draw values of the population proportions from their distribution and calculate the weights using these draws, so that this uncertainty is reflected in the MI variance estimation. An extension of the univariate simulation study discussed above was conducted to address the effect that this extra uncertainty in estimating the incomplete variable's population distribution might have on the simulation results. The extended simulation study compared three cases in which the population distribution was either invariant, estimated in a large external dataset, or estimated in a smaller external dataset compared to the dataset used for analysis. When the population proportions of the binary covariate were estimated in a small external dataset with a higher level of uncertainty, there was an increase in the empirical and average model standard errors in both marginal and conditional weighted MI, particularly when missingness in the covariate was dependent on the outcome. This also led to an increase in the coverage of 95% CIs. The increase in the average model standard errors was due to an increase in the between-imputation variance component of Rubin's variance estimator. Results from this extended simulation study suggested that the extra uncertainty arising from drawing the population proportions from their distribution and calculating the weights was reflected in Rubin's MI variance estimator.

The last part of chapter 3 explored the inclusion of the proposed univariate weighted MI methods in the MICE algorithm [23] for imputing missing values in more than one incomplete covariate (section 3.6). In particular, multivariate simulation studies of a three-way contingency table with a fully observed binary outcome variable y and two incomplete binary covariates x and z were conducted, where the analysis model was a logistic regression of y on x and z. MICE with marginal and conditional weighted conditional models for x and z (referred to as *marginal weighted MICE* and *conditional weighted MICE*) were compared to standard MICE (with unweighted conditional models) in terms of bias, standard errors, and coverage of 95% CIs under three missingness mechanisms for x and z. Simulation results 'tentatively' showed that marginal weighted MICE was the preferred choice. The method yielded small or no bias in point estimates and correct coverage of 95% CIs when missingness in x depended on x (MNAR) and missingness in z depended either on y (MAR) or z (MNAR). When missingness in each covariate depended on its values and the outcome, both marginal weighted MICE and conditional weighted MICE were biased; conditional weighted MICE produced less bias and maintained relatively high coverage.

As seen in the univariate and multivariate simulation studies, neither marginal nor conditional weighted MI could produce unbiased parameter estimates when missingness in the covariate depended on both its values and the outcome. This finding suggested that these weights might not optimally account for both the incomplete covariate's population marginal distribution and the effects of fully observed variables in the imputation model on the incomplete covariate's distribution. Results in chapter 3 thus provided the motivation for calibrated-δ adjustment MI, which was proposed in chapter 4.

7.1.2 *Calibrated-δ adjustment multiple imputation of a binary covariate when the outcome variable is binary*

Chapter 4 proposed and evaluated the univariate *calibrated-δ adjustment multiple imputation* method as an alternative approach to weighting in MI when the population-level marginal distribution of the incomplete covariate is available. In calibrated-δ adjustment MI, the incomplete covariate's population distribution is utilised to calculate an adjustment in the intercept of the imputation model, in order to tackle bias found in marginal and conditional weighted MI when missingness in the covariate depended on both its values and the outcome variable.

The idea of incorporating the calibrated-δ adjustment in MI was motivated by the δ adjustment MI method proposed by van Buuren et al. [23]. The main difference between the two approaches is that while values of δ are often chosen arbitrarily in van Buuren et al.'s method, the incomplete variable's population distribution is used to derive the appropriate δ in calibrated-δ adjustment MI.

Chapter 4 started with an exploration of the equivalence between weighting and including a δ adjustment in the imputation model for the incomplete covariate x in the 2×2 contingency table (section 4.2). Key findings to be noted from this investigation were as follows. When the incomplete covariate x was MNAR dependent on x, or MNAR dependent on x and the outcome variable y, the covariate–outcome association the imputation model for x was the same in the observed and missing data. This implied that adjusting the intercept of the imputation model for x was sufficient to correct bias introduced by missing data under these two MNAR mechanisms. Further, the correct intercept adjustment was shown to be equal to the value of the log odds ratio of observing x for $x = 1$ compared to $x = 0$ in the selection model for x. When missingness in x depended on x, it was found that the correct intercept adjustment was equal to the log ratio of the two marginal weights for $x = 1$ and $x = 0$, which explained why marginal weighted MI was unbiased under this missingness mechanism. However, when missingness in x depended on both x and y, the correct intercept adjustment was neither equal to the log ratios of the two marginal weights nor conditional weights, which illustrated the bias seen in both marginal and conditional weighted MI under this missingness mechanism.

These findings confirmed that in a 2×2 contingency table where both the outcome variable and the incomplete covariate are binary, appropriately adjusting the intercept of the imputation model sufficiently corrects bias in point estimates introduced by the two MNAR mechanisms considered. The derivation of the calibrated-δ adjustment thus involves using the incomplete variable's population marginal distribution as well as its observed-data distribution and association with other fully observed variables to estimate its distribution in the missing data.

Results of this analytic investigation were further affirmed in a univariate simulation study of a 2×2 contingency table (section 4.3). Under all four missingness mechanisms considered, calibrated-δ adjustment MI provided unbiased estimates of the analysis model's parameters, with comparable empirical and average model standard errors and correct coverage of 95% CIs. Most importantly, bias seen in both marginal and conditional weighted MI when x was MNAR dependent on x and y was alleviated by the correct calibrated-δ adjustment.

Further simulations were also conducted to explore the setting where there is uncertainty in estimating the population distribution of the incomplete covariate x (section 4.3.3). More

specifically, it was assumed that the population distribution was estimated in an external dataset of either larger or smaller sizes compared to the dataset used for analysis. Results seen in calibrated-δ adjustment MI were similar to previous results seen in the weighted MI methods. Bias in point estimates slightly increased when the population distribution was estimated in a small external dataset with high uncertainty. This was accompanied by an increase in both the average model and empirical standard errors. Since there was tiny or no bias in the point estimates, coverage was around the 95% level in all cases.

Lastly, the repeated multivariate simulation study conducted to compare the performance of marginal and conditional weighted MICE and standard MICE in chapter 3 was revisited to explore the use of univariate calibrated-δ adjustment MI in the MICE algorithm for handling missing data in more than one covariate (section 4.4). The analysis model considered in this simulation study was a logistic regression of a fully observed binary outcome variable y conditional on two incomplete binary covariates x and z. Three different missingness mechanisms for x and z were investigated. When x was MNAR dependent on x, and z was either MAR conditional on y or MNAR dependent on z, calibrated-δ adjustment MICE appeared to yield small or no bias in all three parameter estimates, which was similar to the results seen in marginal weighted MICE. Standard errors were also comparable for marginal weighted MICE and calibrated-δ adjustment MICE under these two missingness mechanisms, and coverage of both methods attained the nominal level. When missingness in each of the two covariates depended on its values and the outcome, there was minuscule bias in the estimated log odds ratios in calibrated-δ adjustment MICE, which disappeared when the sample size increased from $n = 1\,000$ to $n = 5\,000$ (appendix B.1). However, the empirical standard errors appeared to be larger than the average model counterparts for the estimated log odds ratios in calibrated-δ adjustment MICE, leading to the coverage of the corresponding parameter estimates to be slightly lower than the nominal level. This discrepancy became smaller with increased sample size, and the reason for this discrepancy was unclear.

Calibrated-δ adjustment MI provides an alternative approach to weighting in MI, in which the incomplete variable's population marginal distribution is incorporated into the imputation process via an offset in the imputation model. The calibrated-δ adjustment is calculated using the population marginal distribution of the incomplete variable as well as its distribution and association with other variables in the observed data. Thus, in the univariate missing data setting considered, the method represented a correct approach for utilising population-level information about the incomplete variable while accounting for the effects of covariates in the imputation model. In the univariate and multivariate simulation studies presented thus far, it was found that while marginal weighted MI(CE) could be valid in certain settings, calibrated-δ adjustment MI(CE) was generally preferred to marginal and conditional weighted MI(CE). However, there was still a concern regarding the average model standard errors being slightly smaller than the empirical counterparts in calibrated-δ adjustment MICE, which can affect coverage of the method. This issue was seen in the repeated multivariate simulation study where each of the two incomplete covariates was MNAR dependent on its values and the outcome.

7.1.3 *Population-calibrated multiple imputation of a binary covariate when the outcome variable is continuous*

In chapters 3 and 4, the univariate population-calibrated MI methods were explored in univariate and multivariate missing data settings where both the outcome variable and the incomplete covariate(s) were binary. Chapter 5 studied the application of marginal and conditional weighted MI and calibrated-δ adjustment MI in a univariate missing data setting where the incomplete covariate was binary as before, but the fully observed outcome variable was continuous.

This chapter presented a univariate simulation study conducted to examine finite-sample properties of marginal and conditional weighted MI and calibrated-δ adjustment MI of an incomplete binary covariate x, when the complete outcome variable y is continuous (section 5.2). For moderate coefficient of determination ($R^2 = 0.2$), point estimates were unbiased in standard MI, conditional weighted MI, and calibrated-δ adjustment MI when x was MAR conditional on y. However, while the average model standard errors were comparable in the three methods, the empirical standard errors in calibrated-δ adjustment MI appeared to be slightly larger than the average model standard errors, leading to a small drop in coverage. When missingness in x depended on x, marginal weighted MI, calibrated-δ adjustment MI, and CRA were unbiased. Empirical and average model standard errors were comparable in the population-calibrated MI methods and were smaller than that in CRA. Coverage of all three methods attained the 95% level. Most importantly, when missingness in x depended on x and y, calibrated-δ adjustment MI was no longer unbiased. The empirical standard errors of the method also appeared to be larger than the average model counterparts, leading to coverage falling slightly below the nominal level. For an increased value of the coefficient of determination ($R^2 = 0.5$, appendix C.1), while bias in point estimates disappeared, the discrepancy between the average model and empirical standard errors in calibrated-δ adjustment MI remained present. This discrepancy was less noticeable for the estimated log odds ratio, and coverage of this parameter was also improved.

Previously in a 2×2 contingency table setting, it was noted that the calibrated-δ adjustment MI method was implemented based on the assumption that the association between the incomplete covariate x and the complete outcome y was the same, whether x was observed or missing. This implied that when missingness in x depended on x and y, adjusting the intercept in the imputation model for x was sufficient to correct bias introduced by this MNAR mechanism. To explore whether bias seen in calibrated-δ adjustment MI under this missingness mechanism when y is continuous could be explained by the violation of this assumption, a logistic regression model for x conditional on y, the response indicator r of x, and their interaction was fitted to the full data (i.e. before values in x were set to missing). It was found that the hypothesis regarding the log odds ratio of the interaction term was rejected at 5% level in 13% of the simulation repetitions. Based on this finding, it might not be plausible to assume that the association between x and y was the same among the observed and missing x. This finding also implied that a second adjustment, or sensitivity parameter, was needed in the log odds ratio of the imputation model for x in addition to the existing intercept adjustment. This empirical exploration was also supplemented by an analytic example based on the Heckman model [24] (section 2.5.2). This example provided a mathematical justification for the presence of the second sensitivity parameter (section 5.3).

Findings in the univariate analytic and simulation studies clarified why adjusting the intercept

of the imputation model for x alone was not sufficient to deal with bias introduced by the inclusion of y in the MNAR mechanism for x. Thus, in this case, knowing the population-level marginal distribution of the incomplete covariate is not enough to correctly recover the second sensitivity parameter for the association between the covariate and the outcome variable in the missing data. The problem became exploring the sensitivity of inference for a range of values of the second sensitivity parameter δ_y. This can be done by eliciting δ_y and using the population distribution of the incomplete covariate to derive δ_0, given each elicited value of δ_y. Further simulations were performed to explore the setting where δ_y was fixed to its full-data (i.e. 'correct') estimate (section 5.4). In this approach, bias previously seen in calibrated-δ adjustment MI when δ_y was assumed to be 0 was now removed by fixing δ_y to its estimate obtained in the full data. However, doing so still did not resolve the mismatch between the empirical and average model standard errors, and the reason for this discrepancy was unclear. As a result of this mismatch, coverage was slightly above or below the 95% level, depending on the values of the coefficient of determination.

7.1.4 Case studies using UK primary care electronic health records

In chapter 6, the application of the proposed population-calibrated MI methods was demonstrated using real-life data from The Health Improvement Network (THIN), a large UK primary care electronic health record database (section 6.2.1). Two case studies were conducted using THIN data to illustrate the application of marginal and conditional weighted MI and calibrated-δ adjustment MI for handling missing values in ethnicity, as well as to compare these methods to standard MI and other simple approaches to missing data that are used in practice.

The first case study, which was a cross-sectional study, aimed to assess the plausibility of the MAR assumption for missing data in ethnicity in THIN (section 6.4). The marginal distribution of ethnicity was estimated after missing values in ethnicity were handled by (i) a CRA, (ii) single imputation of missing values with the White ethnic group, (iii) standard MI assuming MAR, (iv) marginal and (v) conditional weighted MI, and (vi) calibrated-δ adjustment MI. The resulting THIN distribution of ethnicity was then compared to the corresponding population distribution in the 2011 UK census data [111]. It was shown that among the complete records, the proportion of the White ethnic group was over-represented, while the proportions of the non-White groups (Asian, Black, Mixed/Other) were under-represented, assuming that the distribution of ethnicity in THIN should match that in the census. Single imputation of missing ethnicity values with the White ethnic group, which assumed that only White individuals ever failed to have their ethnicity recorded, further over-estimated the proportion of the White ethnic group and underestimated the proportions of the non-White ethnic groups. Standard MI also yielded a post-imputation distribution of ethnicity that did not match that in the census. In standard MI, an imputation model was constructed for ethnicity based on individuals' demographics and health conditions. These included age, sex, a measure of social deprivation, as well as indicators of diseases including heart attack, stroke, type 2 diabetes, chronic kidney disease, sickle cell disease, thalassemia, and schizophrenia. These variables were expected to be associated with ethnicity as well as missingness in ethnicity. Even with the extensive inclusion of disease indicators which were thought to be predictive of ethnicity and/or missingness in ethnicity in the imputation model, standard MI still did not recover the distribution in the census, potentially due to a MNAR mechanism for

missing data in ethnicity. In contrast, by incorporating the census distribution of ethnicity in the imputation process, the post-imputation distribution of ethnicity was matched to the population level in marginal and conditional weighted MI as well as calibrated-δ adjustment MI.

The second case study, also a cross-sectional study, extended the setting considered in the first case study by including ethnicity as a covariate in an analysis model to examine the association between ethnicity and the prevalence of type 2 diabetes diagnoses in THIN in 2013 (section 6.5). The analysis model was a logistic regression model for whether an individual had a record indicative of type 2 diabetes, conditional on the individual's demographics including age, sex, and a measure of social deprivation. The resulting odds ratios (OR) and associated standard errors were compared in (i) a CRA, (ii) single imputation of missing values with the White ethnic group, (iii) standard MI assuming MAR, (iv) marginal and (v) conditional weighed MI, and (vi) calibrated-δ adjustment MI, using the census distribution of ethnicity as the reference.

In CRA, the odds of having a diagnosis of type 2 diabetes increased quite smoothly in older age groups, with the most noticeable changes occurring between age 50–59 years and 70–79 years. Results for sex among the complete records indicated that the odds of having a diagnosis of type 2 diabetes was higher for men compared to women. There was also a smooth increase in the odds of having a diagnosis of type 2 diabetes in more deprived quintiles of the Townsend deprivation score. In CRA, individuals of the Asian ethnic group were found to be around 3.5 times more likely to have a diagnosis of type 2 diabetes compared to White individuals. Similarly, the odds of having a diagnosis of type 2 diabetes was higher for the Black ethnic group compared to the White group. Results under single imputation of missing values with the White ethnic group and standard MI were broadly similar to that in CRA, along with a slight increase in the odds of having a type 2 diabetes diagnosis in women.

Results in the population-calibrated MI methods were generally comparable. The estimated ORs increased in older age groups as seen before, but the odds of having a diagnosis of type 2 diabetes was slightly smaller for older age groups in the population-calibrated MI methods compared to the other methods. Therefore, this led to a slight drop in the estimated ORs for these groups. There was an increase towards 1 in the estimated ORs for women in population-calibrated MI compared to the other methods. Estimated ORs also increased in more deprived quintiles of the Townsend deprivation score. However, the estimated ORs grew more rapidly for higher quintiles of deprivation in the population-calibrated MI methods compared to the other methods. The most substantial difference in the results between population-calibrated MI and the rest of the methods could be seen in the estimated ORs of type 2 diabetes diagnoses for ethnicity. Although the Asian and Black ethnic groups were still associated with higher odds of having a diagnosis of type 2 diabetes, the estimated ORs decreased noticeably for these groups. In particular, there was a drop in the ORs from 3.6 for the Asian group in CRA (OR = 3.59; 95% CI 3.43 to 3.75), single imputation with the White ethnic group (OR = 3.63; 95% CI 3.47 to 3.79), and standard MI (OR = 3.58; 95% CI 3.43 to 3.73), to 2.4 in marginal and conditional weighted MI (OR = 2.45; 95% CI 2.34 to 2.56) and calibrated-δ adjustment MI (OR = 2.35; 95% CI 2.36 to 2.45).

Compared to the most complex missingness mechanism investigated in univariate simulation studies for the population-calibrated MI methods where the incomplete covariate was missing depending on its values and the outcome (missingness model M4, tables 3.2 and 4.1), the

assumed missingness mechanism in this study involved an extra element where missingness in the incomplete covariate also depended on other fully observed covariates. It might be reasonable to assume that since these fully observed covariates are all binary or categorical, results under M4 in the univariate simulation studies can be generalised to this case study. That is, if the above missingness assumption holds and the census statistics provide a relevant population distribution of ethnicity for this THIN sample, calibrated-δ adjustment MI can produce more plausible estimates of association in the analysis model. This is done by calibrating the post-imputation distribution of the incomplete covariate to the correct population distribution, while accounting for the relationships with other variables in the observed data.

The next section highlights the implications of the population-calibrated MI methods developed and evaluated in this thesis.

7.2 IMPLICATIONS

Findings in this thesis carry several methodological and practical implications, which are discussed below.

7.2.1 Methodological implications

Most MI analyses are performed under the MAR assumption in practice. It is possible, although more complex, to perform MI assuming missing data are MNAR. In the missing data literature, there are two general approaches for analysing missing data under the MNAR assumption, the selection model [24, 66] and the pattern-mixture model [64, 65]. MI is particularly well-suited for the pattern-mixture modelling approach, since the distribution of missing data across the different missingness patterns can be intervened directly in the imputation process (section 2.5.1). According to Carpenter and Kenward [46], in comparison with the selection modelling approach, the pattern-mixture model is more readily understood and communicated via graphs.

Due to the theoretical complexity of these methods as well as the lack of practical software and/or the requirement to write code using specific packages such as WinBUGS [68], these methods are not frequently adopted in applied settings. These technical difficulties create a barrier for researchers to explore the sensitivity of results under the MNAR mechanism. As a result, MI analyses are increasingly performed under the MAR mechanism without the consideration for alternative MNAR assumptions [22]. Apart from Carpenter et al. [22], most methods for performing sensitivity analyses exploring departure from the MAR assumption are 'un-anchored'. Clinical expert insights are often drawn on for eliciting values of the sensitivity parameters. Alternatively, the sensitivity of inference to alternative MNAR assumptions is examined across a range of different values for the sensitivity parameters in a 'tipping-point' analysis fashion.

In contrast to such methods, the two population-calibrated MI methods proposed in this thesis offer a way to calibrate the data used for analysis to a relevant population-level external dataset, thereby anchoring MI inference to the population, as the name of the methods suggests. At their core, these methods follow a pattern-mixture approach in which the difference in the incomplete variable's distribution between the observed and missing data is implied and represented by information obtained in external population data. Instead of selecting values for

the sensitivity parameter in an arbitrary manner as implemented in van Buuren et al. [23], the population-calibrated MI methods derive values of the sensitivity parameter based on empirical data containing information about the incomplete variable in the population. Further, it is important to highlight that van Buuren et al.'s δ is chosen *independently* of covariates in the imputation model (as commonly done in practice), which can potentially yield implausible imputed values [23]. On the contrary, calibrated-δ adjustment MI accounts or the observed-data association between variables in the imputation model when deriving the value of the δ adjustment. By calibrating MI analysis to the population, the population-calibrated MI methods make a step further compared to other un-anchored approaches for performing sensitivity analyses, since it is sensible that missing data should be imputed in consistence with the population data. The proposed population-calibrated MI methods thus add to the advancement in the field of missing data, particularly in terms of the availability of MI methods for handling missing data under the MNAR mechanism which calibrate inference to the population.

7.2.2 Applied implications: the analyst's perspective

The proposed population-calibrated MI methods were evaluated from a methodological perspective, with investigations carried out in various univariate and multivariate settings using increasingly complex missingness mechanisms for the incomplete covariate(s). These settings ranged from a MCAR mechanism (most restricted) to a MNAR mechanism dependent on the values of the incomplete covariate and the outcome variable. The calibrated-δ adjustment MI method was generally the preferred method across the scenarios considered, while marginal and conditional weighted MI could remove bias or produce more accurate results compared to standard MI in certain settings. In addition, it was shown that calibrated-δ adjustment MI can produce the same inferences as standard MI when data are MAR, and so can conditional weighted MI in certain settings (e.g. when the outcome and covariate are both binary).

In practice, when the analyst is faced with the problem of data suspected to be MNAR in one or several covariates in the analysis, the following steps are recommended for choosing the appropriate MI methods for handling missing values in the incomplete covariates.

1. Define the full-data analysis model, including the outcome variable and the set of covariates (both fully and partially observed);
2. Perform a CRA, which is valid under the assumption that missingness in the covariates is either independent of both the outcome variable and the covariates (MCAR); or is independent of the outcome, conditional on the covariates (MNAR);
3. Select a plausible set of fully observed variables (including the outcome) that are related to the values and missingess in the covariates, and perform a standard MI analysis under the posited MAR assumption;
4. Carefully consider (a) whether the available external datasets are suitable references for the population-calibrated MI methods; (b) the type of the outcome variable and covariates (e.g. binary, continuous); (c) some plausible MNAR mechanisms and whether such MNAR mechanisms depend on the outcome variable and/or other fully observed covariates;
5. Perform the relevant population-calibrated MI method(s) that is (are) valid under the posited MNAR mechanism and compare the results to standard MI and CRA to examine whether

they produce similar inferences;

6. If the methods lead to different conclusions, report the results in all methods and attempt to provide some plausible explanations for the discrepancies seen.

By considering the above steps, it should become evident that knowledge about the incomplete variable's population marginal distribution does not guarantee that such information can be applied in the same way in all situations. Careful considerations of the missingness mechanisms underlying the missing values are therefore unavoidable and also key to the appropriate use of such knowledge in MI.

7.3 STRENGTHS AND LIMITATIONS

The following sections provide some remarks on the strengths and limitations of the proposed population-calibrated MI methods and the investigations carried out in this thesis.

7.3.1 Strengths

As illustrated throughout this thesis, the two proposed population-calibrated MI methods can improve on standard MI by utilising external population data, and hence correct or reduce bias under general MNAR mechanisms. These methods are MI-based with an underlying pattern-mixture modelling nature, and are therefore straightforward to implement and communicate among researchers who are familiar with the use of MI. Further, the imputation procedures described in this thesis could be implemented fairly automatically using a choice of statistical package that offers software for performing MI analysis. As a part of this PhD, I have written and released software for population-calibrated MI in Stata [44], a widely used package in medical research, in order to enable the implementation of these methods in practice.

In a fully Bayesian approach, missing values in the incomplete covariate are treated as extra parameters in the model which require a prior distribution. Knowledge about the population marginal distribution of the incomplete covariate can therefore be incorporated into the model in the form of an informative prior. However, defining a model for the informative prior might not be simple. This is because similar to the use of weights in weighted MI, the prior only reflects the distribution of the incomplete covariate in the missing data, while the aim is to match the completed-data distribution of the covariate to the population level. Therefore, some form of adjustment accounting for the observed-data distribution is also required when defining the informative prior for the missing data, which might be difficult. Compared to a fully Bayesian procedure, calibrated-δ adjustment MI offers a simple solution for incorporating knowledge about the incomplete covariate's population distribution in the imputation process. The calibrated-δ adjustment can be readily calculated based on the reference distribution and used in the imputation model, and adjustments made in the imputation step are separated from the fitting of the analysis model. Further, the implementation of the calibrated-δ adjustment MI procedure in standard statistical software such as Stata [44] or R [119] is reasonably direct and the method is relatively straightforward to communicate, while fitting Bayesian models requires familiarity with specialised software for Bayesian inference.

An important strength of the proposed population-calibrated MI methods is their flexibility

to be adapted to impute variables in a given dataset whose distributions might be available in some external data. For example, in case studies using UK primary care electronic health records in sections 6.4 and 6.5, the census data were used for imputing ethnicity. Similarly, data from other nationally representative datasets such as the Health Survey for England [120] could also be used to impute missing data in other health indicators that are routinely recorded in primary care, such as smoking status or alcohol consumption. In such cases, it is also important to consider the uncertainty associated with estimating the population distributions of these variables, as was highlighted in simulation studies in sections 3.5 and 4.3.3.

Direct linkage of data from individuals included in the analysis to several datasets containing data from the same individuals is increasingly used to reduce the level of missing data in the analysis dataset and obtain a more complete picture of the information available about the individuals' health. For example, primary care electronic health records containing individuals' ethnicity information can be enriched by direct linkage to hospital records in secondary care via the unique patient identifiers. Data linkage thus allows for the direct determination of missing ethnicity information in individuals whose ethnicity is recorded through their hospital visits in secondary care and not in primary care. However, linkage is not attainable for patients who do not allow their data to be linked and thus choose to opt out [121]. The process of linking observations from different data sources can also involve many complications, including the occurrence of linkage errors [106]. In addition, the use of data linkage for improving the completeness of partially observed variables in the analysis dataset requires the linked datasets to contain the same set of variables, which might not always be possible. In contrast, the population-calibrated MI methods do not rely on a direct linkage between observations in the datasets used for analysis and calibration. Instead, the methods 'match' the datasets in the sense that the analysis dataset is assumed to be representative of the population data, i.e. the two datasets are assumed to originate from roughly the same population. Further, the implementation of the population-calibrated MI methods does not require eliciting expert opinions, which can be subjective in nature. Researchers may be more comfortable with utilising objective external empirical data, which is what the proposed population-calibrated MI methods are designed to do.

7.3.2 Limitations

First, it should be noted that not all possible settings can be covered in analytic and empirical investigations. In particular, chapters 3–5 were linked together by a common theme, in which the proposed population-calibrated MI methods were evaluated and compared to existing methods under missingness mechanisms of increasing realism. The work presented in these chapters demonstrated the strengths and limitations of different methods in different scenarios, and the investigations carried out so far relied heavily on simulation studies of various complexity. Although simulation is a useful tool for comparing methods directly since the true data generating mechanisms are known, findings are limited to the scope of the simulation studies considered and does not guarantee that such scenarios are applicable in practice. Therefore, it may be desirable to design simulation studies based on a real-life motivating dataset. However, in this approach, conclusions can also be limited to the nature of the data used to motivate the simulation studies. Nevertheless, it is worth highlighting that analytic and simulation studies conducted in this thesis

represent an attempt to cover practically interesting missingness mechanisms for the missing data that are straightforward to study and interpret.

Second, the implementation of the population-calibrated MI methods proposed in this thesis rests on the availability of relevant external data sources to be used as references in MI. For datasets comprising broad samples of individuals, such as large primary care electronic health records in populations like the UK where the majority of individuals are registered with general practices, there may be a number of population-level external data sources that can be used in MI to calibrate inference to the population (e.g. the UK census data [122] or the Health Survey for England [120]). For other datasets containing information about very specific groups of individuals, such as data from a survey designed to study the experience of patients diagnosed with cancers who are treated in public hospitals, it can be rare or impossible to find external datasets that correspond to the study sample at the population level. Further, it can also happen that although a suitable source of external data can be identified, the data are of relatively poor standard with potential misclassifications and/or missing data. Therefore, depending on the availability and quality of the suitable external data sources, the implementation of the population-calibrated MI methods may or may not be feasible.

Given this remark, in the second case study (section 6.5) which examined the association between ethnicity and the prevalence of type 2 diabetes diagnoses in 2013, the 2011 UK census data were used as reference in the imputation of missing values in ethnicity. This was done based on the assumption that the population composition of the ethnic breakdown did not change very much between 2011 and 2013, which is relatively reasonable. If there is a wider time gap between the analysis and external datasets, results need to be interpreted subject to consideration regarding the representation of the analysis data.

Further, the proposed population-calibrated MI methods were examined in situations where only knowledge about the incomplete variable's population marginal distribution was available. In some settings, such knowledge might not be enough to correct bias introduced by the MNAR mechanism. One such situation was described in sections 5.2 and 5.3, where the continuous outcome variable induced the presence of a second sensitivity parameter for the covariate–outcome association. Another situation is when the MNAR mechanism involves an interaction between the variables, where knowledge of the conditional distributions (i.e. lower level information instead of only marginal) may be needed to remove bias introduced by data being MNAR. Again, the successful implementation of population-calibrated MI depends on whether the necessary information is accessible.

7.4 Remarks on specific findings and further work

From the above discussion regarding the findings and limitations of the investigations carried out in this thesis, several areas for further work and potential extensions are identified below.

7.4.1 Application for more complex analysis models

In the application of the population-calibrated MI methods for handling missing ethnicity data in chapter 6, it was also conjectured that results of the 2×2 contingency table could be generalised

to the case of a three-way or higher-order contingency table with additional fully observed covariate(s) that are binary/categorical. This generalisability was based on the properties of the odds ratios in the (multinomial logistic) regression imputation and analysis models. To confirm this, a simulation study can be conducted featuring a three-way table with a fully observed binary outcome variable, a partially observed categorical covariate, and another fully observed categorical covariate. The analysis model is a logistic regression model for the outcome conditional on the two covariates, and the imputation model for the incomplete covariate is a multinomial logistic regression of the incomplete covariate conditional on the outcome and the other fully observed covariate.

Likewise, further extensions can explore situations where the fully observed covariates are a mixture of continuous and binary/categorical variables, or when the analysis model includes interaction terms between the incomplete and fully observed variables.

7.4.2 Application for incomplete covariates and outcome variables of different types

Motivated by the issue of missing data in ethnicity in UK primary care databases, the development of the population-calibrated MI methods thus far focused on imputing missing values in incomplete binary/categorical covariates. For an incomplete continuous variable whose population marginal distribution (e.g. mean and standard deviation) is available externally, it is less clear how to incorporate such information in MI. This can be explored in further extensions of the population-calibrated MI methods.

It might also be of interest to extend the application of the population-calibrated MI methods in survival analysis. For a time-to-event outcome variable, it is suspected that a second sensitivity parameter for the covariate–outcome association is needed, such as in the case of the continuous outcome considered in chapter 5. Settings involving survival models can be investigated further.

7.4.3 Complexity of the missingness mechanisms

In the multivariate simulation studies presented in sections 3.6.4 and 4.4.3, three missingness mechanisms for x and z were considered over repeated simulations. These ranged from the case where one covariate was MNAR and the other covariate was MAR, to cases where each of the two covariates was MNAR dependent on either its values or both its values and the outcome variable. While these missingness mechanisms did not represent the full set of mechanisms involving three variables in this setting, they were chosen to aid the interpretation of results.

Indeed, in practice the actual missingness mechanisms can be much more complicated, especially when missing values occur in several variables. For example, in the above three-way contingency table setting, missingness in a covariate can depend on its values, the other incomplete covariate which can either be MAR or MNAR, the outcome variable, and/or two-way and three-way interactions. Although more realistic, such complicated missingness mechanisms are much harder to comprehend and they also make it harder to understand the simulation results. Nevertheless, more complex missingness mechanisms can be explored in further simulations.

In the aforementioned univariate simulation studies comparing calibrated-δ adjustment MI and marginal and conditional weighted MI to standard MI and CRA, four missingness mechanisms were considered for the covariate x when the outcome variable y was fully observed.

The complexity of these mechanisms ranged from the case where missingness in x depended on (i) neither y nor x (MCAR, most restricted), either y (MAR conditional on y) or x (MNAR dependent on the covariate), or a sum of the two variables (MNAR dependent on y and x, least restricted). Under these mechanisms, it was shown that using the incomplete covariate's population marginal distribution to calculate the calibrated-δ adjustment in the intercept of the imputation model sufficiently removed bias introduced by missing data. Under a saturated selection model for x where missingness in x depends on both y and x as well as an interaction between the two, it will generally be necessary to adjust the intercept of the imputation model for each combination of x and y (i.e. there will be more than one sensitivity parameter). Hence, under such MNAR mechanisms, knowledge about the incomplete covariate's population marginal distribution alone might not be enough to correct bias introduced by missing data.

7.4.4 Pending issues regarding the standard errors

In the repeated multivariate simulation study presented in section 4.4.3, a possible explanation for the mismatch between the empirical and average model standard errors (from using Rubin's variance estimator) might arise from comparing the calibrated-δ adjustment MICE algorithm to a fully Bayesian approach. In calibrated-δ adjustment MICE, for each incomplete covariate, a univariate conditional imputation model is fitted to subjects with observed values of the covariate to obtain maximum likelihood estimates of the imputation model's parameters, and the δ adjustment is calculated based on these estimates. New parameter values are then drawn from the posterior distribution conditional on the observed data and the δ adjustment to obtain imputed values for the covariate. In a fully Bayesian approach, it might be that for each incomplete covariate, values of the imputation model's parameters are first drawn from the posterior distribution of the parameters conditional on the observed data, followed by calculating values of δ given these draws. This difference in the step for obtaining the calibrated-δ adjustment between the two approaches might be the cause for the discrepancy between the empirical and average model standard errors.

Further investigations can involve updating the current Stata code for implementing the calibrated-δ adjustment MICE algorithm to change the order of deriving the calibrated-δ adjustment in the algorithm, and conducting further simulations to examine the standard errors.

Similarly, in chapter 5, further work is also warranted to gain a better understanding of the discrepancy between the empirical and average model standard errors in calibrated-δ adjustment MI, when the outcome variable was continuous and the second sensitivity parameter was fixed to its estimate in the full data.

7.5 CONCLUSION

The proposed population-calibrated MI methods, including marginal and conditional weighted MI and calibrated-δ adjustment MI, represent pragmatic and practical approaches for utilising external population information about the incomplete variable(s) in the imputation process. These methods offer a formal way for researchers to incorporate information obtained from external data sources in MI, in order to assess the plausibility of the MAR assumption in the

analysis of incomplete data. By investigating missing data scenarios that are realistic and relevant in practice, the work conducted in this thesis demonstrated that these methods are likely to perform well and can potentially lead to more accurate inferences compared to standard MI under the MNAR mechanism.

By matching the analysis dataset to external population data, the proposed population-calibrated MI methods anchor inference to the population level, and therefore provide practical tools for performing sensitivity analyses to potential departure from the MAR assumption. At the very least, findings from this thesis highlighted the importance of considering the plausibility of the MAR assumption in the presence of missing data. Researchers should therefore be encouraged to perform sensitivity analyses under alternative MNAR assumptions using all available information, and consider results from such analyses in companion with the results obtained in standard MI assuming data are MAR.

References

[1] Hardoon S, Hayes JF, Blackburn R, Petersen I, Walters K, Nazareth I, Osborn DPJ. Recording of severe mental illness in United Kingdom primary care, 2000-2010. *PLoS ONE*, 8(12): e82365, 2013.

[2] Osborn DP, Hardoon S, Omar RZ, Holt RI, King M, Larsen J, Marston L, Morris RW, Nazareth I, Walters K, Petersen I. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. *JAMA Psychiatry*, 72(2):143–151, 2015. doi: 10.1001/jamapsychiatry.2014.2133.

[3] Man SL, Petersen I, Thompson M, Nazareth I. Antiepileptic drugs during pregnancy in primary care: a UK population based study. *PloS ONE*, 7(12):e52339, 2012. doi: 10.1371/journal.pone.0052339.

[4] Petersen I, McCrea RL, Osborn DJP, Evans S, Pinfold V, Cowen PJ, Gilbert R, Nazareth I. Discontinuation of antipsychotic medication in pregnancy: a cohort study. *Schizophrenia Research*, 159(1):218–225, 2014. doi: 10.1016/j.schres.2014.07.034.

[5] McCrea RL, Nazareth I, Evans SJW, Osborn DPJ, Pinfold V, Cowen PJ, Petersen I. Lithium prescribing during pregnancy: a UK primary care database study. *PLoS ONE*, 10(3): e0121024, 2015. doi: 10.1371/journal.pone.0121024.

[6] Wijlaars LPMM, Nazareth I, Petersen I. Trends in depression and antidepressant prescribing in children and adolescents: a cohort study in the health improvement network (THIN). *PLoS ONE*, 7(3):e33181, 2012. doi: 10.1371/journal.pone.0033181.

[7] Rait G, Walters K, Bottomley C, Petersen I, Iliffe S, Nazareth I. Survival of people with clinical diagnosis of dementia in primary care: cohort study. *BMJ*, 341:c3584, 2010. doi: 10.1136/bmj.c3584.

[8] Grant RL, Drennan VM, Rait G, Petersen I, Iliffe S. First diagnosis and management of incontinence in older people with and without dementia in primary care: a cohort study using The Health Improvement Network primary care database. *PLoS Medicine*, 10(8): e1001505, 2013. doi: 10.1371/journal.pmed.1001505.

[9] Bresnahan M, Begg MD, Brown A, Schaefer C, Sohler N, Insel B, Vella L, Susser E. Race and risk of schizophrenia in a US birth cohort: another example of health disparity? *International Journal of Epidemiology*, 36:751–758, 2007. doi: 10.1093/ije/dym041.

[10] Scarborough P, Bhatnagar P, Kaur A, Smolina K, Wickramasinghe K, Rayner M. Ethnic differences in cardiovascular disease: 2010 edition. Report, Deaprtment of Public Health, University of Oxford, 2010. URL https://www.bhf.org.uk/publications/statistics/ethnic-differences-in-cardiovascular-disease-2010.

[11] Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Current Diabetes Report*, 13(6):1–18, 2013. doi: 10.1007/s11892-013-0421-9.Race/Ethnic.

[12] Mathur R, Grundy E, Smeeth L. Availability and use of UK based ethnicity data for health research. *National Centre for Research Methods Working Paper Series*, 2013. URL http://eprints.ncrm.ac.uk/3040/.

[13] Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, VanStaa T, Grundy E, Smeeth L. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *Journal of Public Health*, pages 1–9, 2013. doi: 10.1093/pubmed/fdt116.

[14] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, 335:136, 2007. doi: 10.1136/bmj.39261.471806.55.

[15] Kumarapeli P, Stepaniuk R, De Lusignan S, Williams R, Rowlands G. Ethnicity recording in general practice computer systems. *Journal of Public Health*, 28(3):283–287, 2006. doi: 10.1093/jpubhealth/fdl044.

[16] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264, 1995.

[17] Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. doi: 10.1037/1082-989X.7.2.147.

[18] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338:b2393, 2009. doi: 10.1136/bmj.b2393.

[19] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336:a332, 2008. doi: 10.1136/bmj.39609.449676.25.

[20] Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987.

[21] Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999.

[22] 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*, 23:1352–1371, 2013. doi: 10.1080/10543406.2013.834911.

[23] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681–694, 1999.

[24] Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimators for such models. *Annals of Economic and Social Measurement*, 5(4):475–492, 1976.

[25] Morris TP. *Practical use of multiple imputation*. PhD thesis, University College London, 2014. URL http://discovery.ucl.ac.uk/1419272/3/Morris_Timothy_Thesis_amended.pdf.

[26] Carpenter J, Plewis I. Analysing longitudinal studies with non-response: issues and statistical methods. In *The SAGE Handbook of Innovation in Social Research Methods*, chapter 23. 2011.

[27] Little RJA, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New Jersey, 2002. doi: 10.1002/9781119013563.

[28] Rubin DB. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

[29] Little RJA. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

[30] 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*, 1:368–376, 2004. doi: 10.1191/1740774504cn032oa.

[31] Rezvan PH, 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*, 15:30, 2015. doi: 10.1186/s12874-015-0022-1.

[32] Graham JW. Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009. doi: 10.1146/annurev.psych.58.110405.085530.

[33] Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, 134, 1991.

[34] Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91(433):222– 230, 1996.

[35] Cook RJ, Zeng L, Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, 60(3):820–828, 2004. doi: 10.1111/j.0006-341X.2004.00234.x.

[36] Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Journal of American Statistics Association*, 61(1):79–90, 2007.

[37] van Buuren S. *Flexible imputation of missing data*. Chapman & Hall/CRC, Boca Raton, 2012.

[38] Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992. doi: 10.2307/2290664.

[39] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29:2920–2931, 2010. doi: 10.1002/sim.3944.

[40] Bartlett JW, Carpenter JR, Tilling K, Vansteelandt S. Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics*, 15(4):719–730, 2014. doi: 10.1093/biostatistics/kxu023.

[41] Smuk M. *Missing data methodology: sensitivity analysis after multiple imputation*. PhD thesis, London School of Hygiene and Tropical Medicine, 2015. URL http://researchonline.lshtm.ac.uk/2212896/.

[42] Yuan Y. Multiple imputation using SAS software. *Journal of Statistical Software*, 45(6), 2011. doi: 10.18637/jss.v045.i06.

[43] van Buuren S, Groothuis-Oudshoorn K. `mice`: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011. doi: 10.1177/0962280206074463.

[44] StataCorp. *Stata statistical software: release 14*. StataCorp LP., College Station, TX, 2015. College Station, TX: StataCorp LP.

[45] Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996. doi: 10.1080/01621459.1996.10476908.

[46] Carpenter JR, Kenward MG. *Multiple imputation and its application*. John Wiley & Sons, Ltd., Chichester, West Sussex, 1st edition, 2013.

[47] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30:377–399, 2011. doi: 10.1002/sim.4067.

[48] Schafer JL. *Analysis of incomplete multivariate data*. Chapman & Hall/CRC, London, 1997.

[49] Carlin JB. Multiple imputation: a perspective and historical overview. In *Handbook of missing data methodology*, chapter 12, pages 239–266. Chapman & Hall/CRC, 2015.

[50] Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15, 1999. doi: 10.1177/096228029900800102.

[51] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007. doi: 10.1177/0962280206074463.

[52] Tanner M, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[53] van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006. doi: 10.1080/10629360600810434.

[54] Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JAC. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14:28, 2014.

[55] Liu J, Gelman A, Hill J, Su YS, Kropko J. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014. doi: 10.1093/biomet/ast044.

[56] von Hippel PT. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39:265–291, 2009.

[57] Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59:1092–1101, 2006. doi: 10.1016/j.jclinepi.2006.01.009.

[58] White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28:1982–1998, 2009. doi: 10.1002/sim.3618.

[59] Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35, 2003. doi: 10.1111/1467-9574.00218.

[60] Meng XL. Inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994. doi: 10.2307/2246252.

[61] Robins JM, Wang N. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 2000. doi: 10.2307/2673565.

[62] 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*, 12:46, 2012. doi: 10.1186/1471-2288-12-46.

[63] Morris TP, White IR, Royston P, Seaman SR, Wood AM. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine*, 33:88–104, 2014. doi: 10.1002/sim.5935.

[64] Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. doi: 10.2307/2290705.

[65] Little RJA. A class of pattern-mixture models for multivariate incomplete data. *Biometrika*, 81(3):471–483, 1994.

[66] Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16: 259–275, 2007. doi: 10.1177/0962280206075303.

[67] Puhani PA. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1):53–68, 2000. doi: 10.1111/1467-6419.00104.

[68] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000. doi: 10.1023/A:1008929526011.

[69] White IR, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials*, 4:125–139, 2007. doi: 10.1177/1740774507077849.

[70] Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Statistics in Medicine*, 21(8):1043–1066, 2002. doi: 10.1002/sim.1065.

[71] Hayati Rezvan P, White IR, Lee KJ, Carlin JB, Simpson JA. Evaluation of a weighting approach for performing sensitivity analysis after multiple imputation. *BMC Medical Research Methodology*, 15:83, 2015. doi: 10.1186/s12874-015-0074-2.

[72] Raghunathan T. *Missing data analysis in practice*. Chapman & Hall/CRC, Boca Raton, 2015.

[73] Pham TM. *MI_IMPUTE_WLOGIT: Stata module to perform weighted multiple imputation for binary/categorical variables*. Statistical Software Components. Boston College Department of Economics, Boston, 2016. URL https://ideas.repec.org/c/boc/bocode/s458241.html.

[74] Royston P. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

[75] StataCorp LP. *Stata multiple-imputation reference manual*. Stata Press, College Station, TX, 2015. URL http://www.stata.com/manuals14/mi.pdf.

[76] Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9(1):56, 2009. doi: 10.1186/1471-2288-9-56.

[77] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. 25:4279–4292, 2006. doi: 10.1002/sim.2673.

[78] White IR. simsum: Analyses of simulation studies including Monte Carlo error. *The Stata Journal*, 10(3):369–385, 2010.

[79] Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology*, 182(8):730–736, 2015. doi: 10.1093/aje/kwv114.

[80] Leacy FP. *Multiple imputation under missing not at random assumptions via fully conditional specification*. PhD thesis, University of Cambridge, 2016.

[81] Russ SB. A translation of Bolzano's paper on the Intermediate Value Theorem. *Historia Mathematica*, 7(2):156–185, 1980. doi: 10.1016/0315-0860(80)90036-1.

[82] Burden RL, Faires JD. *Numerical Analysis*. Brooks/Cole, Cengage Learning, Boston, 9th edition, 2010.

[83] Medicines & Healthcare Products Regulatory Agency. Welcome to the Clinical Practice Research Datalink, . URL https://www.cprd.com/intro.asp.

[84] IMS Health Real World Evidence Solutions. THIN data. URL http://www.epic-uk.org/.

[85] QRESEARCH. QRESEARCH specialises in research & analyses using primary care electronic health data. URL http://www.qresearch.org/.

[86] Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Informatics in Primary Care*, 19:251–255, 2011.

[87] Chisholm J. The Read clinical classification. *BMJ*, 300(6732):1092, 1990.

[88] Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiology and Drug Safety*, 18:704–707, 2009. doi: 10.1002/pds.1770.

[89] Joint Formulary Committee. *British National Formulary (BNF) 72*. BMJ Publishing Group Ltd & Royal Pharmaceutical Society, 72nd edition, 2016.

[90] Townsend P. *Health and deprivation: inequality and the north*. Croom Helm, London, 1988.

[91] Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiology and Drug Safety*, 18:76–83, 2009. doi: 10.1002/pds.1688.

[92] Horsfall L, Walters K, Petersen I. Identifying periods of acceptable computer usage in primary care research databases. *Pharmacoepidemiology and Drug Safety*, 22:64–69, 2013. doi: 10.1002/pds.3368.

[93] Shephard E, Stapley S, Hamilton W. The use of electronic databases in primary care research. *Family Practice*, 28:352–354, 2011. doi: 10.1093/fampra/cmr039.

[94] QRESEARCH. A summary of public health indicators using electronic data from primary care. Report, The NHS Information Centre for Health and Social Care, 2008. URL http://content.digital.nhs.uk/catalogue/PUB04632/publ-heal-indi-data-prim-care-01-07-rep.pdf.

[95] Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiology and Drug Safety*, 19:618–626, 2010. doi: 10.1002/pds.1934.

[96] Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, White IR, Petersen I. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study. *BMJ Open*, 4:e004958, 2014. doi: 10.1136/bmjopen-2014- 004958.

[97] Welch C. *Implementation, evaluation and application of multiple imputation for missing data in longitudinal electronic health record research*. PhD thesis, University College London, 2015. URL http://discovery.ucl.ac.uk/1464072/1/Catherine.Welch.thesis_final%5B1%5D.pdf.REDACTED.pdf.

[98] Scott T, Maynard A. Will the new GP contract lead to cost effective medical practice? Report, Centre for Health Economics, University of York, 1991. URL http://www.york.ac.uk/media/che/documents/papers/discussionpapers/CHE Discussion Paper 82.pdf.

[99] BMA. QOF guidance. URL https://www.bma.org.uk/advice/employment/contracts/gp-partner-contracts/qof-guidance.

[100] The NHS Information Centre for Health and Social Care. Quality and Outcomes Framework achievement data 2008/09. Report, 2009. URL http://content.digital.nhs.uk/catalogue/PUB04033/qof-eng-08-09-bull-rep.pdf.

[101] Delaney JAC, Moodie EEM, Suissa S. Validating the effects of drug treatment on blood pressure in the General Practice Research Database. *Pharmacoepidemiology and Drug Safety*, 17:535–545, 2008. doi: 10.1002/pds.1553.

[102] Hull SA, Mathur R, Badrick E, Robson J, Boomla K. Recording ethnicity in primary care: assessing the methods and impact. *British Journal of General Practice*, pages e290–e294, 2011. doi: 10.3399/bjgp11X572544.

[103] Aspinall PJ, Jacobson B. Why poor quality of ethnicity data should not preclude its use for identifying disparities in health and healthcare. *Quality and Safety in Health Care*, 16: 176–180, 2007. doi: 10.1136/qshc.2006.019059.

[104] IMS Health. Data content. URL http://www.epic-uk.org/our-data/data-content.shtml.

[105] Medicines & Healthcare Products Regulatory Agency. CPRD linked data, . URL https://www.cprd.com/dataAccess/linkeddata.asp.

[106] Hagger-Johnson G, Harron K, Fleming T, Gilbert R, Goldstein H, Landy R, Parslow RC. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open*, 5(8):e008118, 2015. doi: 10.1136/bmjopen-2015-008118.

[107] Nitsch D, Kadalayil L, Mangtani P, Steenkamp R, Ansell D, Tomson C, Dos Santos Silva I, Roderick P. Validation and utility of a computerized South Asian names and group recognition algorithm in ascertaining South Asian ethnicity in the national renal registry. *QJM: An International Journal of Medicine*, 102(12):865–872, 2009. doi: 10.1093/qjmed/hcp142.

[108] Ryan R, Vernon S, Lawrence G, Wilson S. Use of name recognition software, census data and multiple imputation to predict missing data on ethnicity: application to cancer registry records. *BMC Medical Informatics and Decision Making*, 12(3), 2012. doi: 10.1186/1472-6947-12-3.

[109] Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open*, 3:e003389, 2013. doi: 10.1136/bmjopen-2013-003389.

[110] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001. doi: 10.1037/1082-989X.6.4.330.

[111] Office for National Statistics. Ethnicity and national identity in England and Wales: 2011. Report, 2012. URL https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicityandnationalidentityinenglandandwales/2012-12-11.

[112] Office for National Statistics. Ethnic group. Report, 2015. URL http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/ethnic-group.pdf.

[113] StataCorp LP. *Stata user's guide*. Stata Press, College Station, TX, 2015. URL https://www.stata.com/manuals14/u.pdf.

[114] Spratt M, Carpenter J, Sterne JAC, Carlin JB, Heron J, Henderson J, Tilling K. Strategies for multiple imputation in longitudinal studies. *American Journal of Epidemiology*, 172(4):478–487, 2010. doi: 10.1093/aje/kwq137.

[115] Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clinical Epidemiology*, 8:373–380, 2016. doi: 10.2147/CLEP.S113415.

[116] Sharma M, Nazareth I, Petersen I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. *BMJ Open*, 6(1):e010210, 2016. doi: 10.1136/bmjopen-2015-010210.

[117] Kenward MG, Carpenter JR. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16:199–218, 2007. doi: 10.1177/0962280206075304.

[118] Rezvan PH, 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*, 15:30, 2015. doi: 10.1186/s12874-015-0022-1.

[119] R Development Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org.

[120] UK Data Service. Health Survey for England. URL https://discover.ukdataservice.ac.uk/series/?sn=2000021.

[121] NHS Digital. Care information choices. URL https://data.gov.uk/dataset/care-information-choices/resource/0f715c81-788b-46a3-a732-89773cf8cdf4.

[122] Office for National Statistics. Welcome to the Office for National Statistics. URL https://www.ons.gov.uk/.

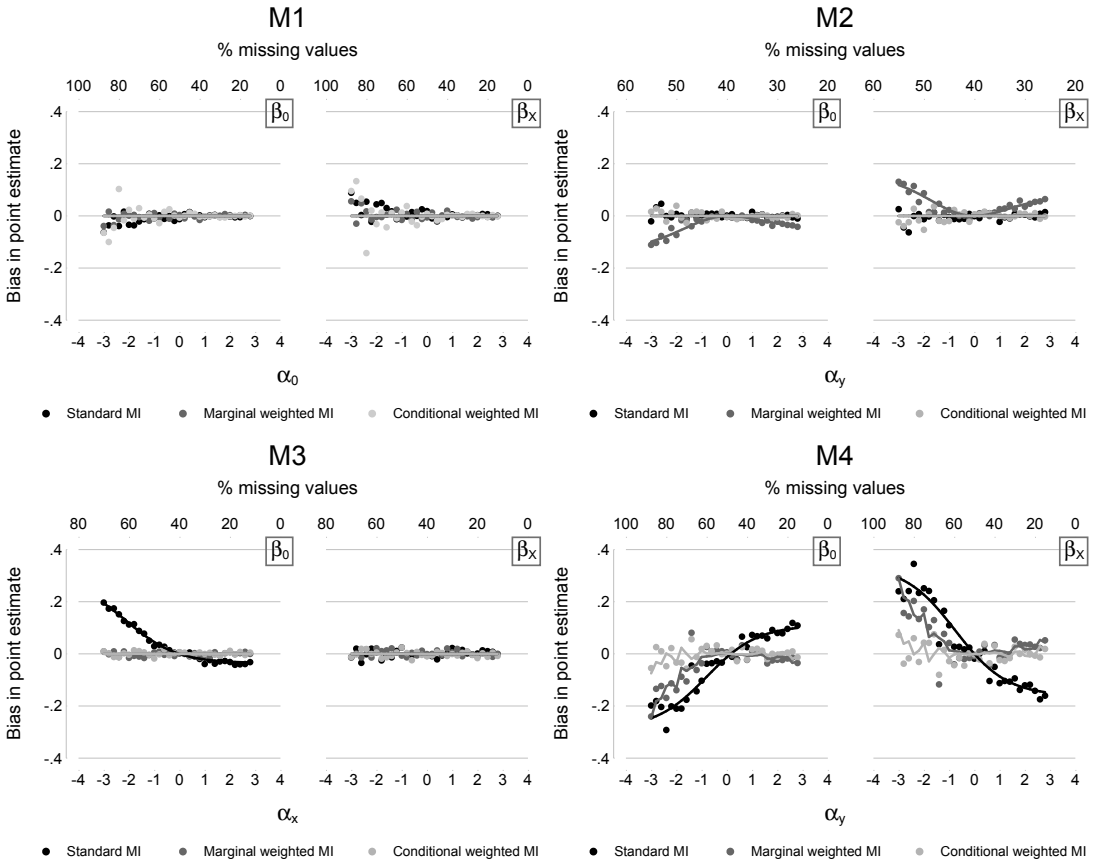
A

Supplementary materials for chapter 3

A.1 VERIFICATION OF ANALYTIC CALCULATIONS USING SIMULATION

As discussed in section 3.3.2, figure A.1 depicts the results of initial two-dimensional simulations which are performed to verify analytic bias calculations in a 2×2 contingency table under various missingness mechanisms considered for the incomplete covariate x.

Figure A.1. Analytic study: comparison of bias in point estimates obtained analytically and empirically via simulation under different missingness mechanisms for x.



* Note: circles: analytic bias; lines: empirical bias; M1: missingness in x does not depend on x or y, $\alpha_o = [-3, 3]$; M2: missingness in x depends on y, $\alpha_o = 0.5$, $\alpha_y = [-3, 3]$; M3: missingness in x depends on x, $\alpha_o = 0.5$, $\alpha_x = [-3, 3]$; M4: missingness in x depends on (x, y), $\alpha_o = \alpha_x = 0.5$, $\alpha_y = [-3, 3]$; bias is plotted against the corresponding percentages of missing values in x.

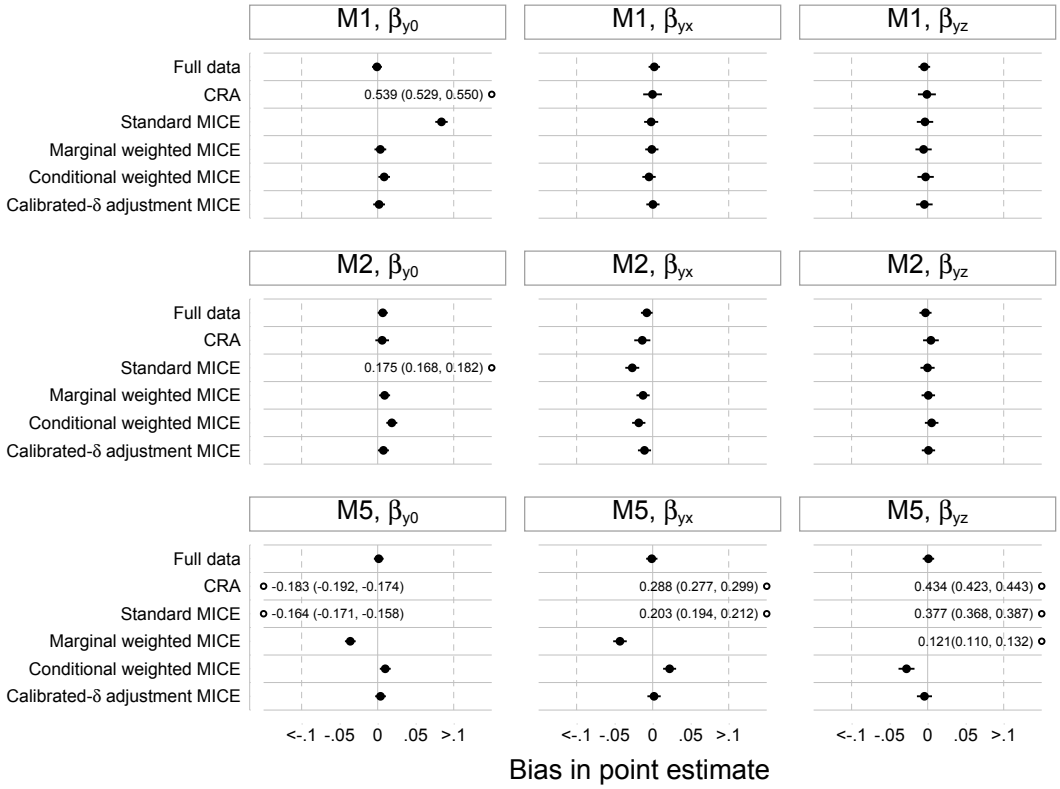
B

Supplementary materials for chapter 4

B.1 REPEATED SIMULATIONS FOR ASSESSING PERFORMANCE MEASURES

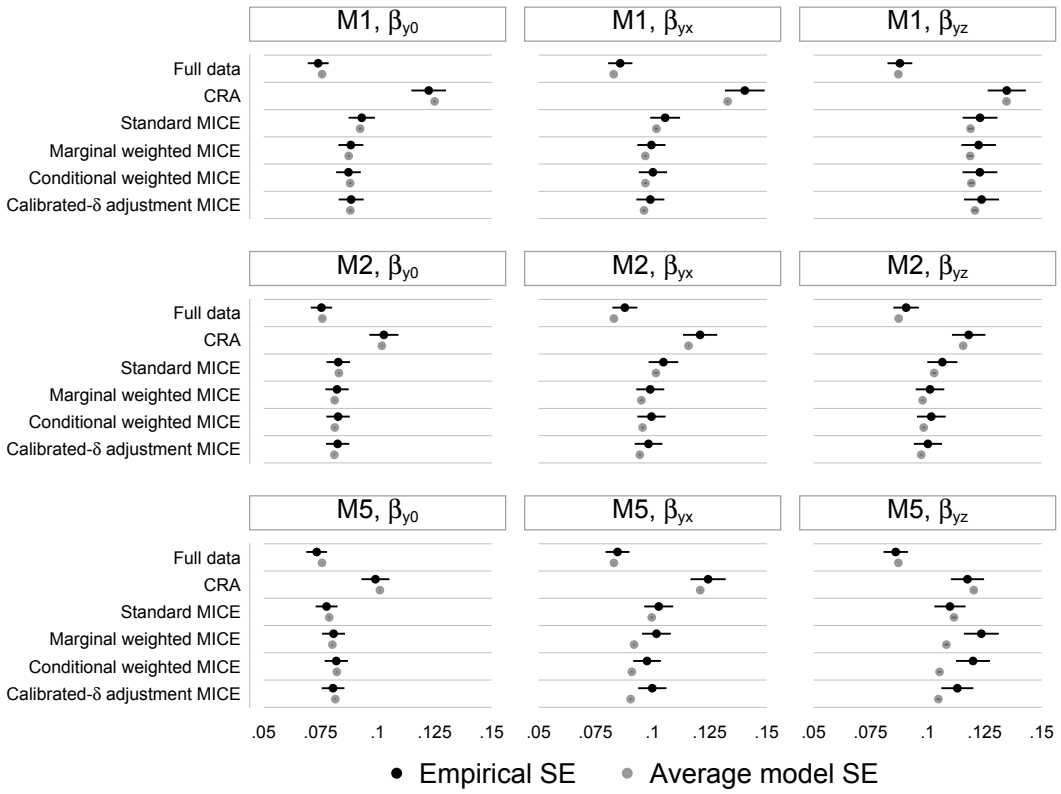
Figures B.1–B.6 present the results of the repeated multivariate simulation study discussed in section 4.4.3, for sample sizes n = 3 000 and 5 000.

Figure B.1. Repeated multivariate simulation study (n = 3 000): bias in point estimates under different missingness mechanisms for x and z.



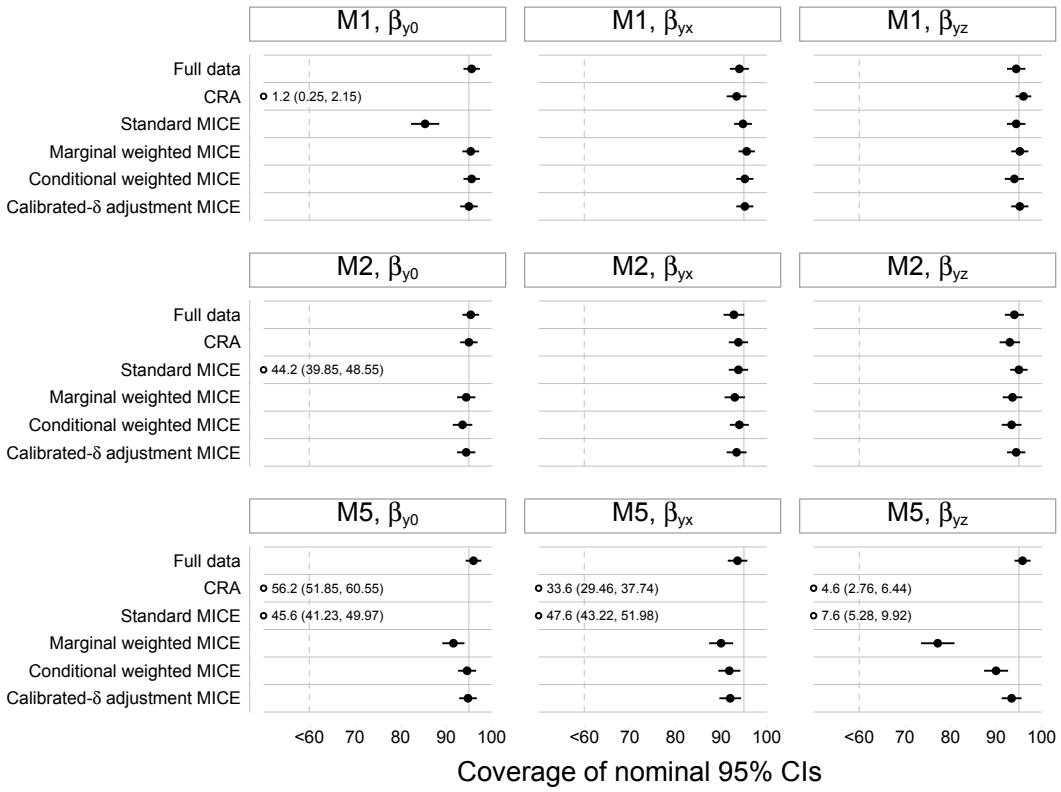
⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); $\beta_0 = 0.5, \beta_x = -1, \beta_z = 1$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure B.2. Repeated multivariate simulation study (n = 3 000): empirical and average model standard errors under different missingness mechanisms for x and z.



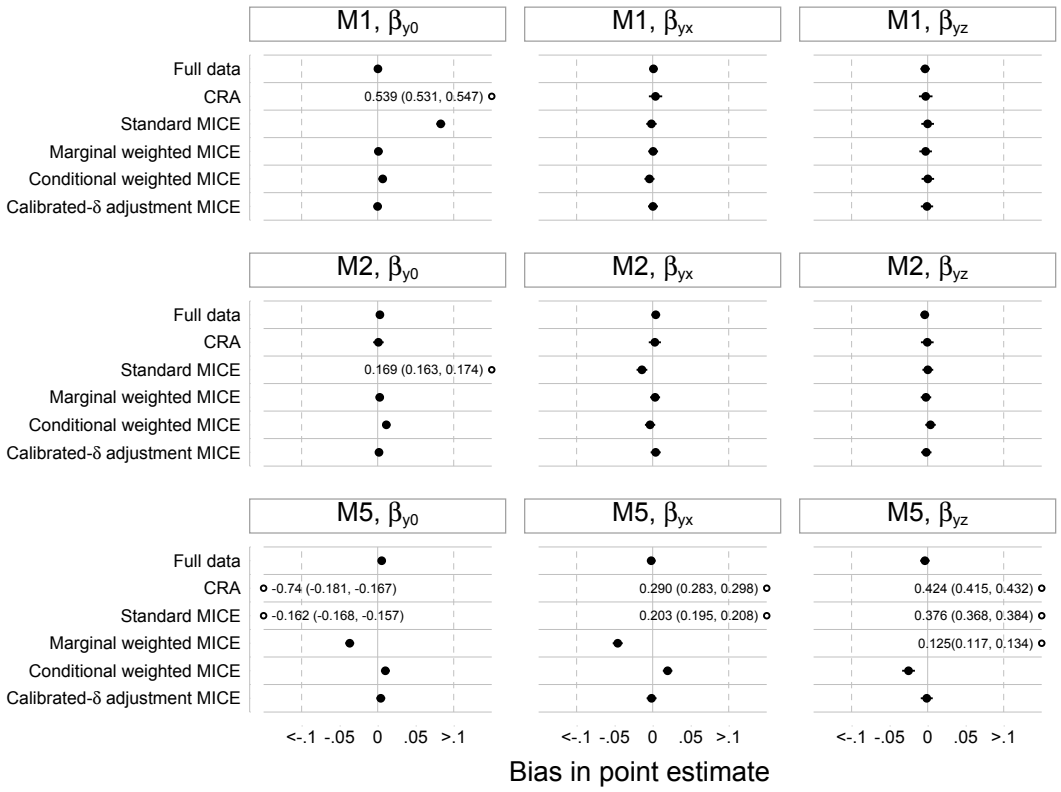
⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors.

Figure B.3. Repeated multivariate simulation study ($n = 3\,000$): coverage of nominal 95% confidence intervals under different missingness mechanisms for x and z.



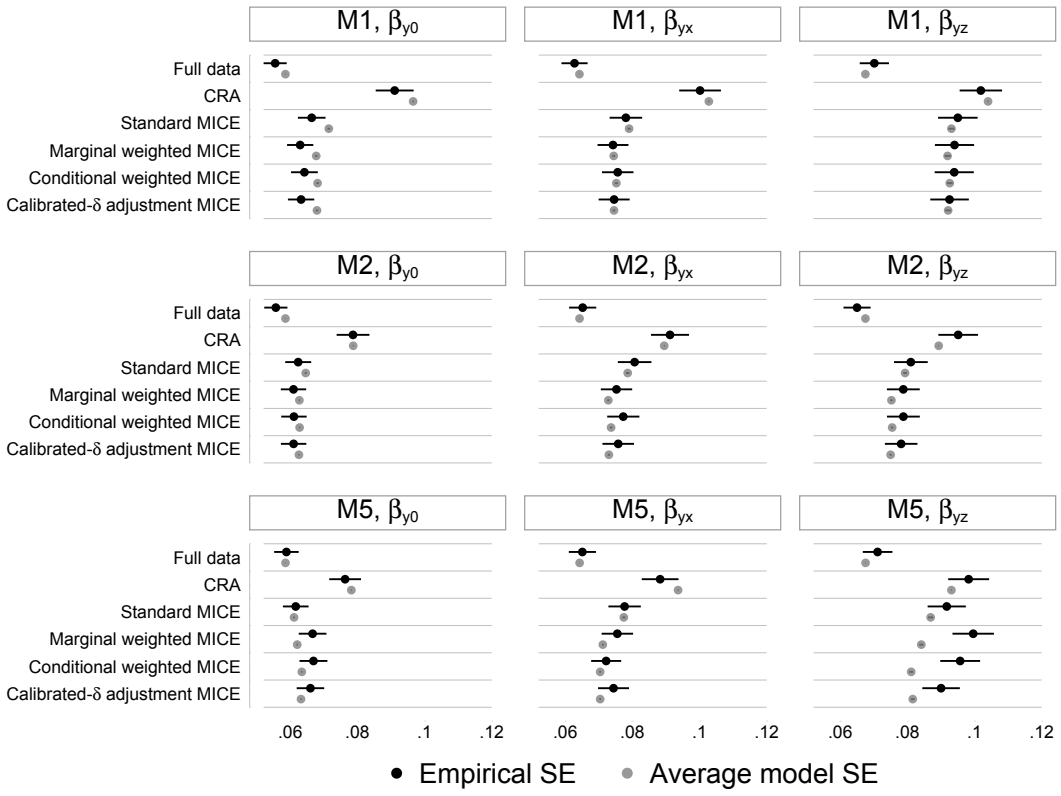
⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure B.4. Repeated multivariate simulation study (n = 5 000): bias in point estimates under different missingness mechanisms for x and z.



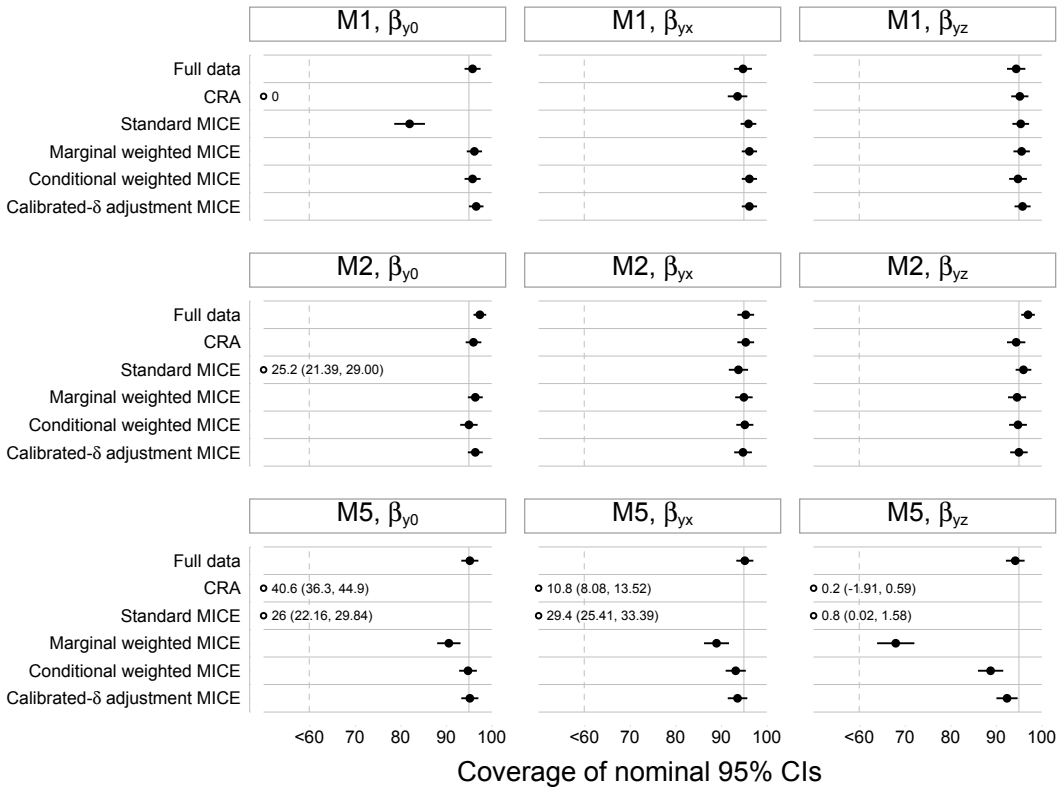
⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); $\beta_o = 0.5, \beta_x = -1, \beta_z = 1$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure B.5. Repeated multivariate simulation study (n = 5 000): empirical and average model standard errors under different missingness mechanisms for x and z.



⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors.

Figure B.6. Repeated multivariate simulation study ($n = 5\,000$): coverage of nominal 95% confidence intervals under different missingness mechanisms for x and z.



⋆ Note: M1: missingness in x depends on x and in z depends on y; M2: missingness in x depends on x and in z depends on z; M5: missingness in x depends on (x, y) and in z depends on (z, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

C

Supplementary materials for chapter 5

C.1 UNIVARIATE SIMULATION STUDY

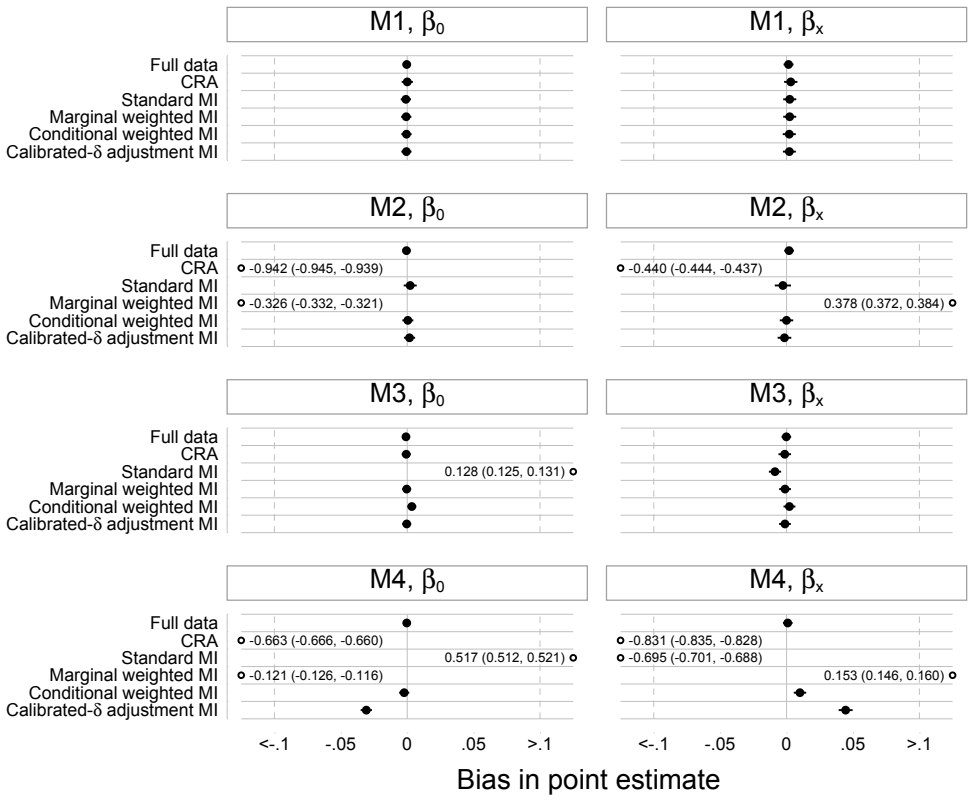
This section presents the results of the univariate simulation study discussed in section 5.2 for different values of R^2.

Figures C.1–C.3 summarise the results of the univariate simulation study when calibrated-δ adjustment MI is implemented with one sensitivity parameter, assuming the association between x and y is the same in the observed and missing data. Under M2, when $R^2 = 0.05$, the empirical standard errors of both conditional weighted MI and calibrated-δ adjustment MI are smaller than the average model counterparts, leading to an over-coverage of 95% CIs. When $R^2 = 0.5$, the empirical and average model standard errors of conditional weighted MI are now comparable. However, the empirical standard errors of calibrated-δ adjustment MI are larger than the average model standard errors, particularly for $\hat{\beta}_0$, which corresponds to a drop in coverage. Under M4, when $R^2 = 0.05$, bias in calibrated-δ adjustment MI is noticeable, while the method appears unbiased when $R^2 = 0.5$. Empirical standard errors are larger than the average model standard errors in calibrated-δ adjustment MI, and coverage slightly decreases when $R^2 = 0.5$; these results are similar to that when $R^2 = 0.2$.

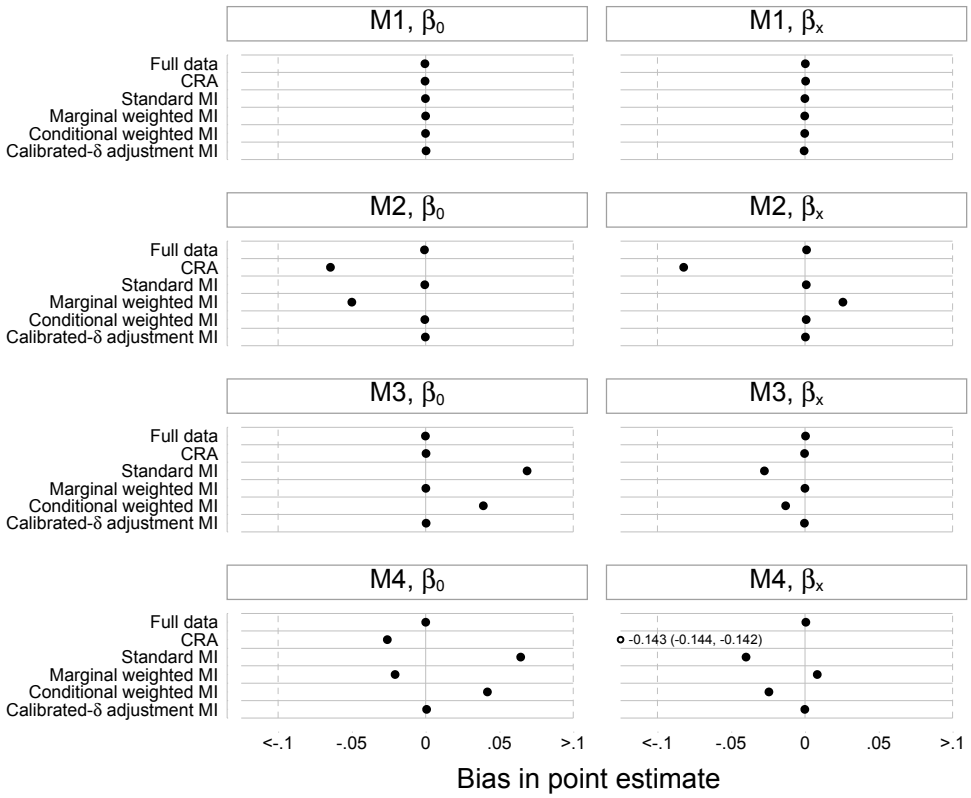
Table C.1 shows the mean and standard deviation (SD) of the estimates of θ_r and θ_{yr} over $S = 1000$ simulation repetitions, and the number of times each of the null hypotheses $H_0 : \theta_r = 0$ and $H_0 : \theta_{yr} = 0$ is rejected at 5% level. There is an increase in the standard deviation of θ_{yr} when $R^2 = 0.5$ compared to when $R^2 = 0.05$. The number of times the hypothesis concerning θ_{yr} is rejected drops from nearly 40% of the simulation repetitions when $R^2 = 0.05$ to 5% when $R^2 = 0.5$, which explains the decrease in bias seen in calibrated-δ adjustment MI assuming $\delta_y = 0$.

Figure C.1. Univariate simulation study (R^2 = 0.05 and 0.5): bias in point estimates under different missingness mechanisms for x.

a. R^2 = 0.05



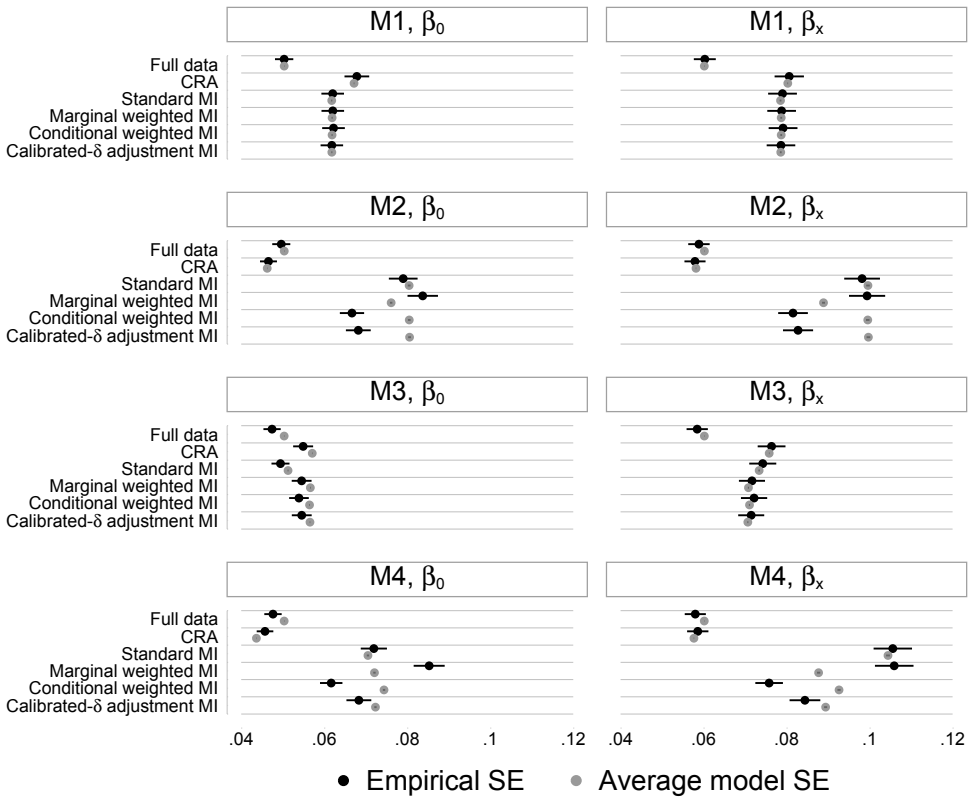
b. R^2 = 0.5



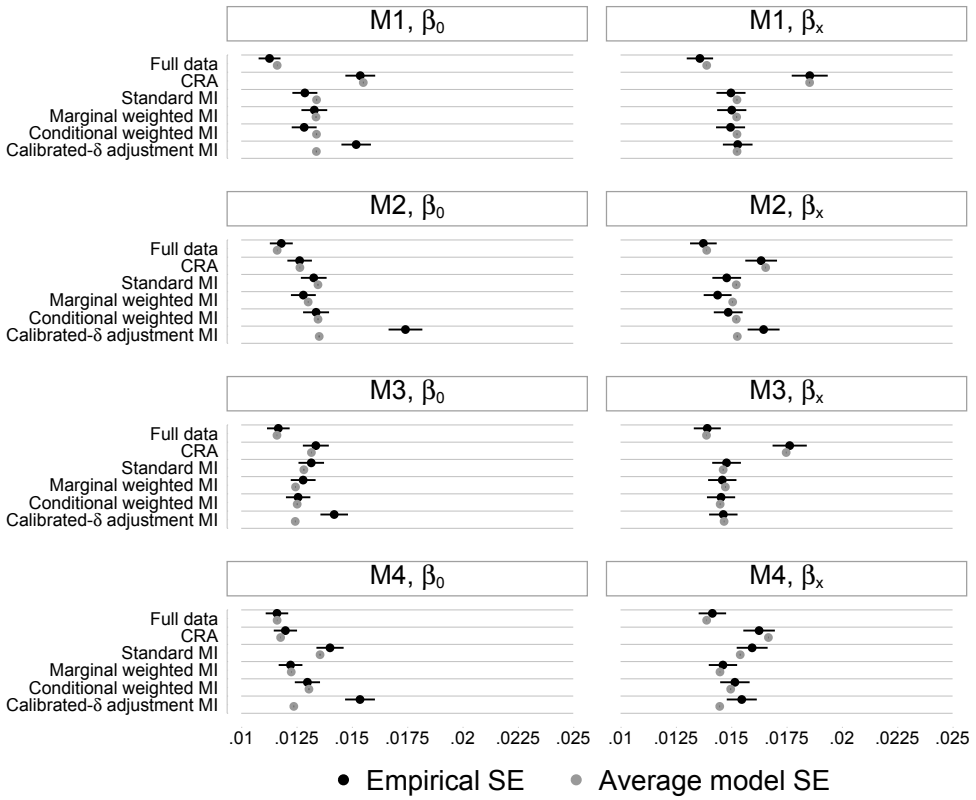
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); $\beta_0 = -0.5, \beta_x = 1$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure C.2. Univariate simulation study (R^2 = 0.05 and 0.5): empirical and average model standard errors under different missingness mechanisms for x.

a. R^2 = 0.05



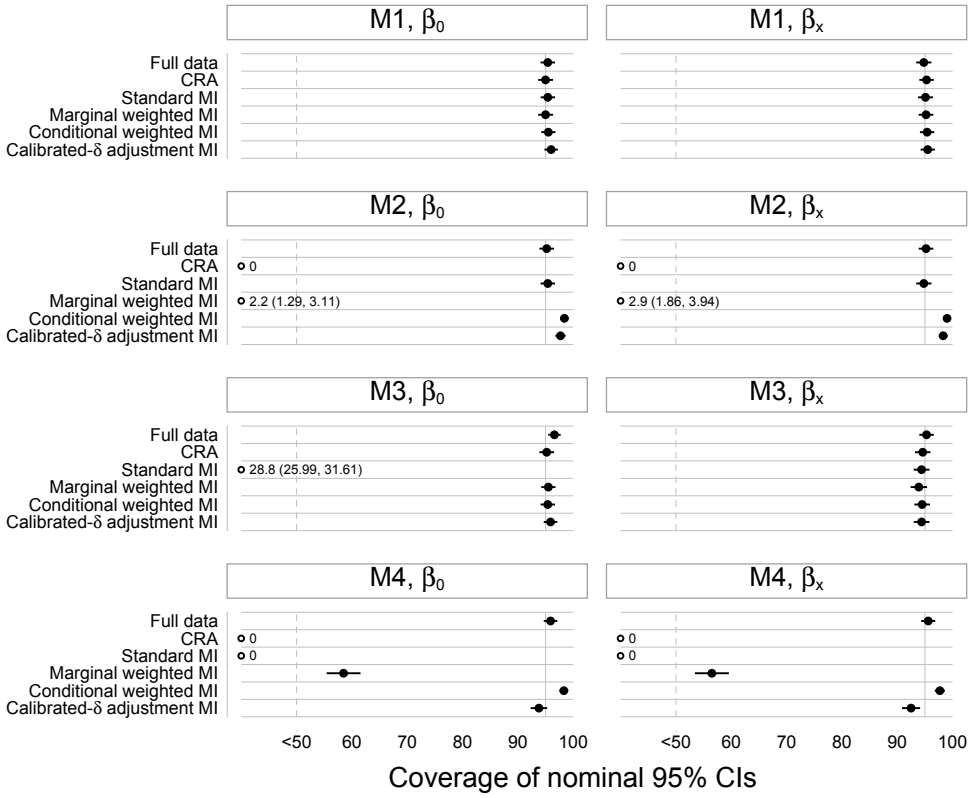
b. R^2 = 0.5



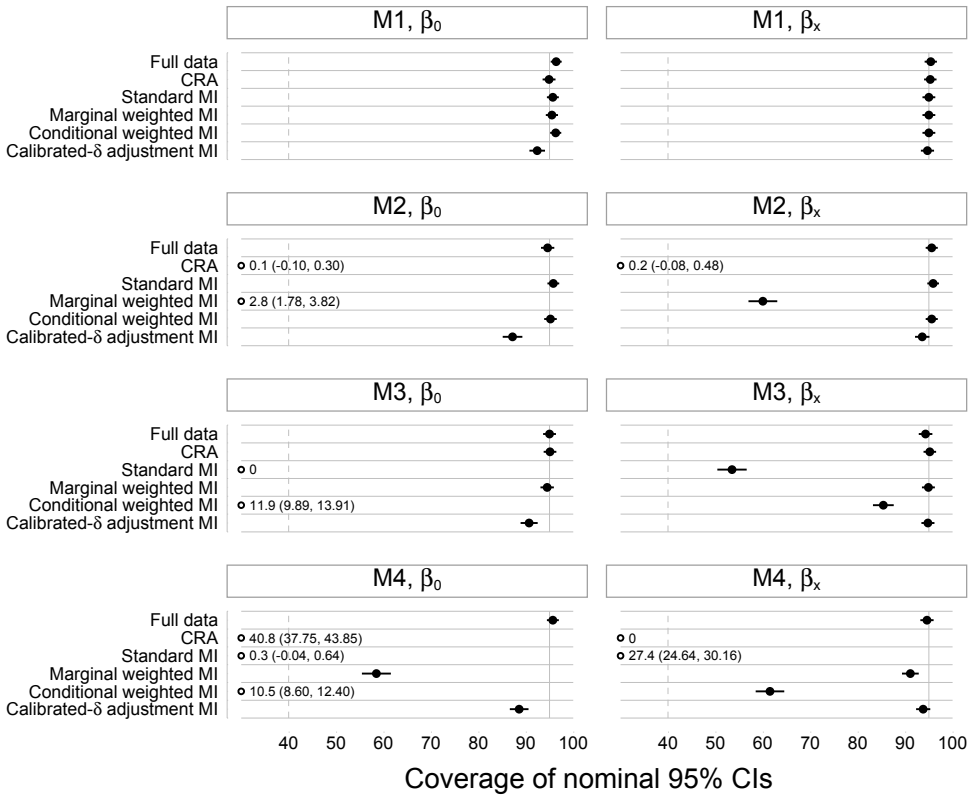
* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors.

Figure C.3. Univariate simulation study (R^2 = 0.05 and 0.5): coverage of nominal 95% confidence intervals under different missingness mechanisms for x.

a. R^2 = 0.05



b. R^2 = 0.5



* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y); error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Table C.1. Univariate simulation study (R^2 = 0.05 and 0.5): mean and standard deviation (SD) of the full-data estimates of θ_r and θ_{yr} over S = 1 000 simulation repetitions and number of times each of the null hypotheses $H_0 : \theta_r = 0$ and $H_0 : \theta_{yr} = 0$ is rejected at the 5% level.

a. R^2 = 0.05

Missingness model	$\bar{\hat{\theta}}_r$	SD $\left(\hat{\theta}_r\right)$	$\bar{\hat{\theta}}_{yr}$	SD $\left(\hat{\theta}_{yr}\right)$	Number of times $H_0 : \theta_r = 0$ rejected	Number of times $H_0 : \theta_{yr} = 0$ rejected
M1	0.0004	0.0634	0.0006	0.0329	51	43
M2	−0.0034	0.0891	−0.0004	0.0484	49	63
M3	−1.5014	0.0759	−0.0007	0.0390	1000	51
M4	−1.3620	0.0984	0.0828	0.0461	1000	391

b. R^2 = 0.5

Missingness model	$\bar{\hat{\theta}}_r$	SD $\left(\hat{\theta}_r\right)$	$\bar{\hat{\theta}}_{yr}$	SD $\left(\hat{\theta}_{yr}\right)$	Number of times $H_0 : \theta_r = 0$ rejected	Number of times $H_0 : \theta_{yr} = 0$ rejected
M1	−0.0041	0.0959	−0.0102	0.2941	44	47
M2	−0.0008	0.1058	−0.0094	0.3068	53	44
M3	−1.5074	0.1127	−0.0174	0.3123	1000	44
M4	−1.5037	0.1247	0.0092	0.3536	1000	52

* Note: M1: missingness in x does not depend on x or y; M2: missingness in x depends on y; M3: missingness in x depends on x; M4: missingness in x depends on (x, y).

C.2 THEORETICAL JUSTIFICATION OF THE ADDITIONAL SENSITIVITY PARAMETER

This section describes the simulation performed to verify the calculations shown in section 5.3.

C.2.1 Method

A single large simulation is conducted; if the calculations are not correct, the discrepancy between the empirical and theoretical results given by the calculations will be apparent and detectable in a large simulated dataset.

The data generating mechanism and analysis procedures are as follows.

1. Simulate n = 1 000 000 complete values of the binary covariate x from the Bernoulli distribution, $x \sim \text{Bernoulli}\left(p_x^{\text{pop}}\right) = 0.7$;
2. Simulate data for the continuous, normally distributed outcome y from the linear regression model for y conditional on x, such that

$$y = \beta_0 + \beta_x x + \varepsilon_{y|x};$$
$$\varepsilon_{y|x} \sim \text{N}\left(0, \sigma^2_{y|x}\right),$$

   where values of 0.5, 2, and 1 are arbitrarily chosen for β_0, β_x, and $\sigma_{y|x}$, respectively;
3. Simulate data for the (latent) continuous, normally distributed variable z which governs the missingness in x from the linear regression model for z conditional on x and y

$$z = \alpha_0 + \alpha_x x + \alpha_y y + \varepsilon_{z|x,y};$$
$$\varepsilon_{z|x,y} \sim \text{N}(0, 1),$$

   where α_0, α_x, and α_y are arbitrarily set to 2, 0.5, and −0.5, respectively;

4. Simulate a binary indicator of response r of x, such that $r = 1$ if $z \geq 0$, and $r = 0$ otherwise;
5. Fit the logistic regression imputation model for x conditional on y

$$\text{logit}\left[p\left(x = 1 \mid y\right)\right] = \theta_0 + \theta_y y,$$

to the observed (i.e. $r = 0$) and missing (i.e. $r = 1$) data in turn, recording the parameter estimates $\hat{\theta}_0$ and $\hat{\theta}_y$ in the observed and missing data;
6. Compare the empirical results to that given by the following calculations

$$\theta_0^{\text{obs}} = \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x\right)}{\Phi\left(\alpha_0\right)}\right] - \frac{2\beta_0\beta_x + \beta_x^2}{2\sigma_{y|x}^2} + \ln\left(\frac{p_x}{1 - p_x}\right);$$

$$\theta_y^{\text{obs}} = \ln\left[\frac{\Phi\left(\alpha_0 + \alpha_x + \alpha_y\right)\Phi\left(\alpha_0\right)}{\Phi\left(\alpha_0 + \alpha_y\right)\Phi\left(\alpha_0 + \alpha_x\right)}\right] + \frac{\beta_x}{\sigma_{y|x}^2};$$

$$\theta_0^{\text{mis}} = \ln\left[\frac{1 - \Phi\left(\alpha_0 + \alpha_x\right)}{1 - \Phi\left(\alpha_0\right)}\right] - \frac{2\beta_0\beta_x + \beta_x^2}{2\sigma_{y|x}^2} + \ln\left(\frac{p_x}{1 - p_x}\right);$$

$$\theta_y^{\text{mis}} = \ln\left\{\frac{\left[1 - \Phi\left(\alpha_0 + \alpha_x + \alpha_y\right)\right]\left[1 - \Phi\left(\alpha_0\right)\right]}{\left[1 - \Phi\left(\alpha_0 + \alpha_y\right)\right]\left[1 - \Phi\left(\alpha_0 + \alpha_x\right)\right]}\right\} + \frac{\beta_x}{\sigma_{y|x}^2}.$$

C.2.2 Results

Table C.3 shows a comparison of the imputation model's parameters in the observed and missing data. These are obtained empirically and analytically by following the calculations presented in section 5.3. All percentage differences between the empirical and theoretical results are small (less than 1%), supporting the validity of the calculations.

The values of the imputation model's intercept in the observed and missing data are noticeably different, suggesting that an intercept adjustment is needed in the imputation model for x when the model is fitted to the observed data. In addition, there is also a difference between the two log odds ratios in the observed and missing data. This finding is consistent with the theoretical results regarding the presence of a second sensitivity parameter for the covariate–outcome association in the imputation model for x. This second sensitivity parameter represents the difference in the association of x and y between the observed and missing data.

Table C.3. Comparison of parameters θ in the imputation model for the covariate x obtained empirically and analytically, when the outcome variable y is continuous.

	θ_0^{obs}	θ_y^{obs}	θ_0^{mis}	θ_y^{mis}
Empirical	−2.1461	2.0493	−3.4457	2.2157
Analytical	−2.1359	2.0294	−3.4512	2.2212
% difference	0.475	0.983	0.157	0.251

C.3 Univariate simulation study: when the second sensitivity parameter is fixed to its full-data estimate

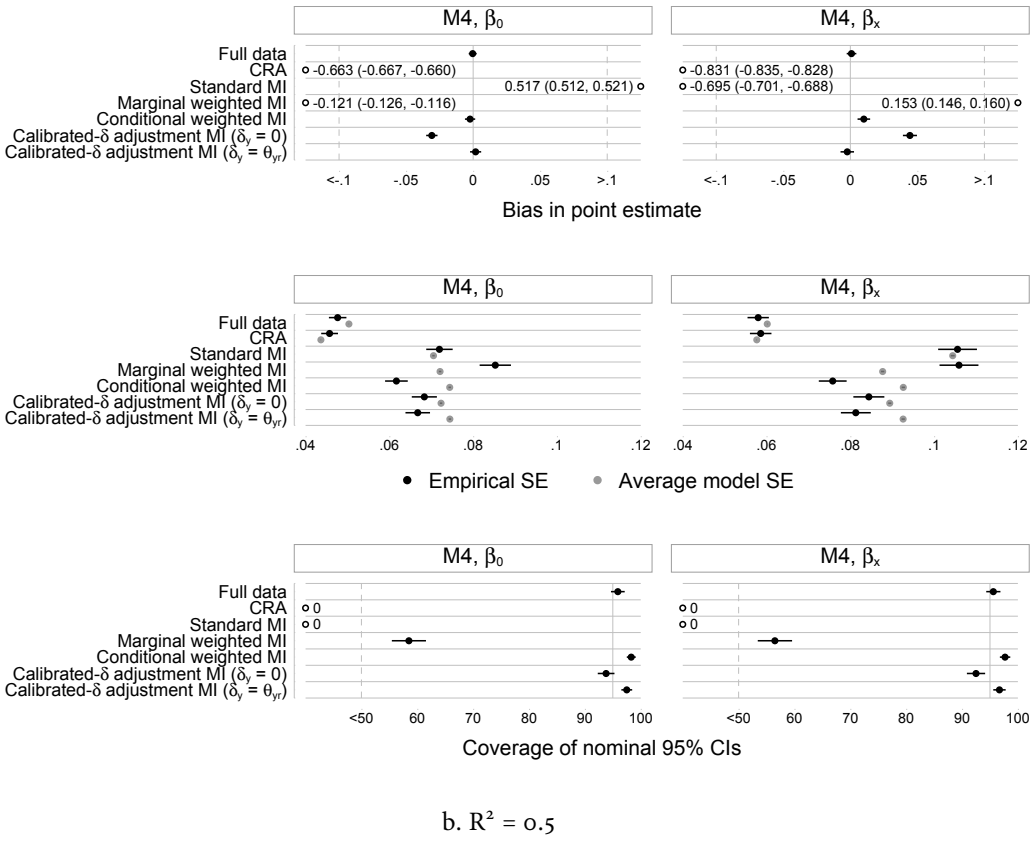
This section presents the results of the univariate simulation study discussed in section 5.4 for different values of R^2.

Figure C.4 summarises the results of the univariate simulation study under M4, when calibrated-δ adjustment MI is implemented with δ_y set to its full-data estimate and δ_0 is derived from the population distribution of x, given the fixed δ_y. This method appears unbiased for both values of R^2. There is still a difference between the empirical and average model standard errors, which corresponds to the slight over- or under-coverage of 95% CIs. This coverage issue is more noticeable for $\hat{\beta}_0$ when $R^2 = 0.5$.

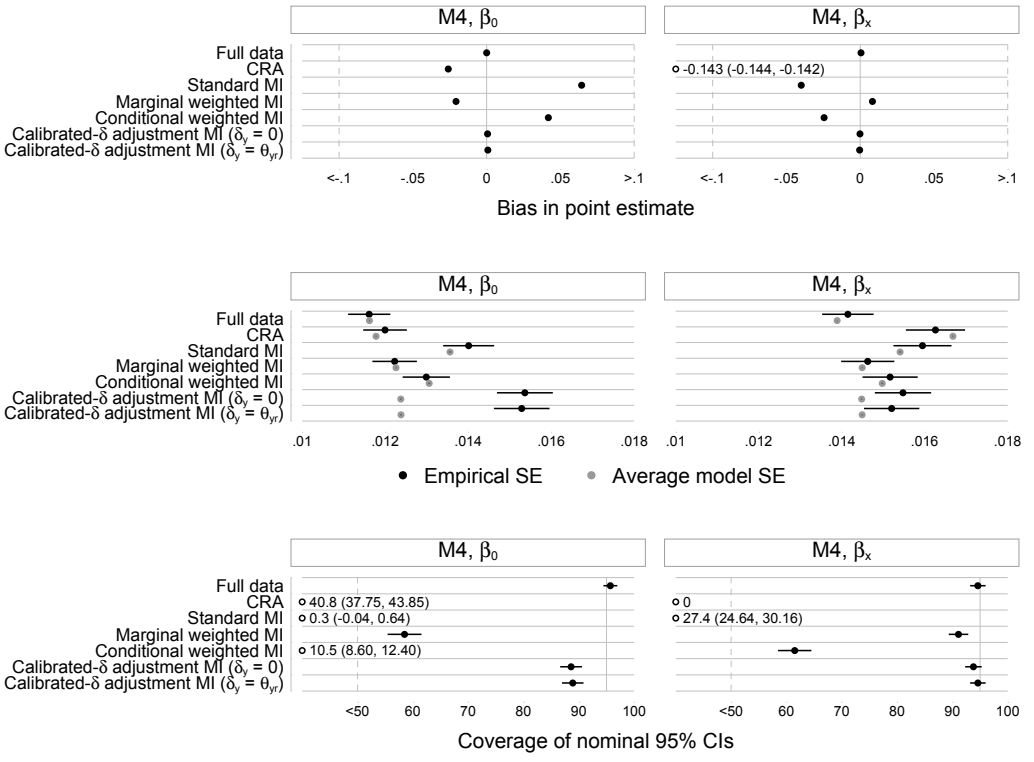
A comparison of $\hat{\theta}_r$ obtained in the full data; calibrated δ_0 derived assuming $\delta_y = \hat{\theta}_{yr}$ where $\hat{\theta}_{yr}$ is estimated in the full data; and calibrated δ_0 derived assuming $\delta_y = 0$ is presented in figure C.5. The difference between the location of the distribution of calibrated δ_0 when $\delta_y = 0$ and the other two quantities is larger when R^2 is small. This again explains bias seen in calibrated-δ adjustment MI with one sensitivity parameter when $R^2 = 0.05$. The spread of the distributions of calibrated δ_0 is wider than that of $\hat{\theta}_r$, which is more noticeable for higher R^2.

Figure C.4. Univariate simulation study (R^2 = 0.05 and 0.5): bias in point estimates, empirical and average model standard errors, and coverage of nominal 95% confidence intervals when missingness in x depends on x and y (M4).

a. R^2 = 0.05

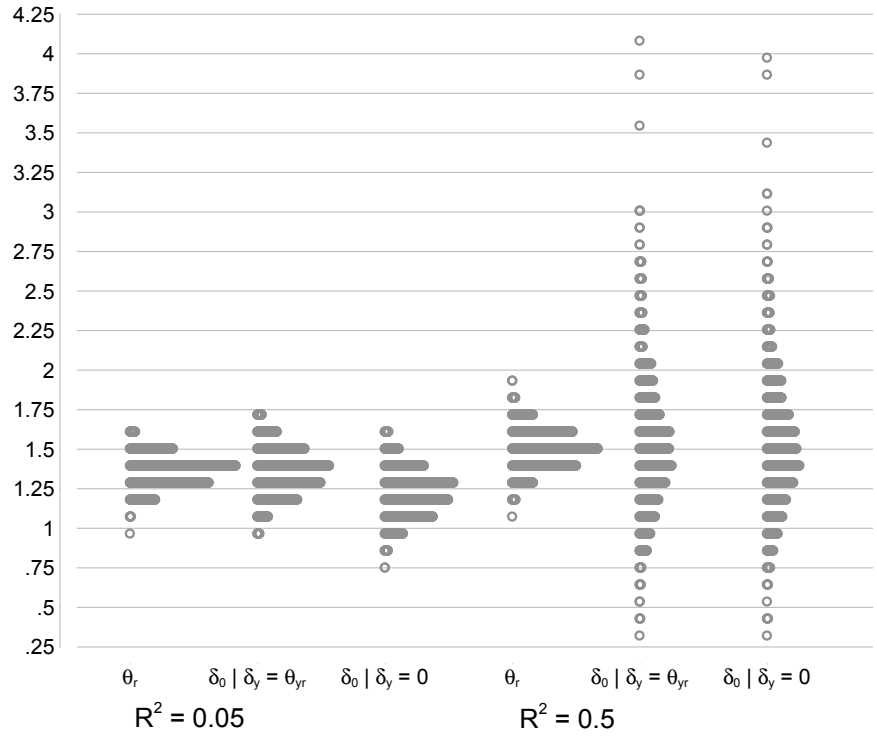


b. R^2 = 0.5



* Note: $\beta_0 = -0.5, \beta_x = 1$; error bars: ±1.96× Monte Carlo standard errors; hollow circles: out-of-range values.

Figure C.5. Univariate simulation study (R^2 = 0.05 and 0.5): comparison of $\hat{\theta}_r$ estimated in the full data; calibrated δ_0 derived assuming $\delta_y = \hat{\theta}_{yr}$, where $\hat{\theta}_{yr}$ is estimated in the full data; and calibrated δ_0 derived assuming $\delta_y = 0$ over $S = 1\,000$ simulation repetitions, when missingness in x depends on x and y (M4).



D

Supplementary materials for chapter 6

D.1 Read codes for extracting ethnicity information in the health improvement network database

A Read code list for ethnicity has been developed by my colleague, Dr Emre Basatemur, which contains 904 codes used in primary care to record information about individuals' ethnicity/-nationality/race, country of birth, language, or requirement for an interpreter. Some codes do not provide enough information to determine individuals' ethnicity, in which case ethnicity is classified as missing. Below are the 50 most frequently used codes for extracting ethnicity information in the two THIN case studies. Altogether these codes accumulate a cumulative frequency of 93%.

Read code	Description
9S10.00	White British
9i0..00	British or Mixed British - ethnic category 2001 census
9S1..00	White
9S13.00	White Scottish
9i00.00	White British - ethnic category 2001 census
9i20.00	English - ethnic category 2001 census
9i2..00	Other White background - ethnic category 2001 census
9iG..00	Ethnic category not stated - 2001 census
9S12.00	Other White ethnic group
13l4.00	Main spoken language English
9i7..00	Indian or British Indian - ethnic category 2001 census
9SE..00	Ethnic group not recorded
9SD..00	Ethnic group not given - patient refused
9iC..00	African - ethnic category 2001 census
9S6..00	Indian
9S3..00	Black African
9i...00	Ethnic category - 2001 census
9i2F.00	Polish - ethnic category 2001 census
9S...00	Ethnic groups (census)
9SJ..00	Other ethnic group
9i8..00	Pakistani or British Pakistani - ethnic category 2001 census

Read code	Description
9i2R.00	Other White European/European unspecified/Mixed European 2001 census
9S7..00	Pakistani
9iA..00	Other Asian background - ethnic category 2001 census
9i22.00	Welsh - ethnic category 2001 census
9i1..00	Irish - ethnic category 2001 census
9S11.00	White Irish
9iB..00	Caribbean - ethnic category 2001 census
9i21.00	Scottish - ethnic category 2001 census
9iE..00	Chinese - ethnic category 2001 census
9i2T.00	Other White or White unspecified ethnic category 2001 census
9S14.00	Other White British ethnic group
9SH..00	Other Asian ethnic group
9iF..00	Other - ethnic category 2001 census
9S2..00	Black Caribbean
226..00	O/E - ethnic group
134B.00	Race: Caucasian
9S9..00	Chinese
9i9..00	Bangladeshi or British Bangladeshi - ethnicity category 2001 census
13lC.00	Main spoken language Polish
134..00	Country of origin
9i6..00	Other Mixed background - ethnic category 2001 census
9i24.00	Northern Irish - ethnic category 2001 census
9i3..00	White and Black Caribbean - ethnic category 2001 census
13Z6.00	Language spoken
13l..00	Main spoken language
134N.00	Race: White
13dC.00	Born in England
9iAA.00	Other Asian or Asian unspecified ethnic category 2001 census
9i4..00	White and Black African - ethnic category 2001 census

D.2 Associations of ethnicity and the response indicator of ethnicity with fully observed variables in case studies 1 and 2

The following tables present unadjusted associations of ethnicity and the response indicator of ethnicity with fully observed variables in case studies 1 and 2 (sections 6.4 and 6.5). The p-values presented here are obtained from performing χ^2 tests of independence in a two-way contingency table for each of the fully observed variables considered.

Table D.2. Case study 1: unadjusted associations of ethnicity with variables used to inform the imputation of ethnicity among the complete records; n = 337 278.

Variable	White		Asian		Black		Mixed/Other		∑		p-value
	Frequency	%	Frequency	%	Frequency	%	Frequency	%	Frequency	%	
Sex											< 0.001
Male	115 982	72.28	19 000	11.84	15 377	9.58	10 114	6.30	160 473	100	
Female	129 082	73.01	18 519	10.47	17 997	10.18	11 207	6.34	176 805	100	
Age group (years)											< 0.001
0–9	25 181	62.87	5 214	13.02	5 605	13.99	4 054	10.12	40 054	100	
10–19	19 578	64.22	3 668	12.03	4 423	14.51	2 815	9.23	30 484	100	
20–29	33 497	69.53	6 590	13.68	4 450	9.24	3 642	7.56	48 179	100	
30–39	43 126	69.69	8 624	13.94	5 907	9.55	4 223	6.82	61 880	100	
40–49	37 989	72.01	5 195	9.85	6 498	12.32	3 073	5.83	52 755	100	
50–59	29 906	76.71	3 651	9.37	3 497	8.97	1 931	4.95	38 985	100	
60–69	25 619	84.28	2 442	8.03	1 397	4.6	941	3.1	30 399	100	
70–79	16 708	83.37	1 607	8.02	1 255	6.26	471	2.35	20 041	100	
80+	13 460	92.82	528	3.64	342	2.36	171	1.18	14 501	100	
Townsend score											< 0.001
Quintile 1 (least deprived)	34 293	87.64	2 417	6.18	1 203	3.07	1 215	3.11	100		
Quintile 2	43 712	84.67	4 009	7.77	1 970	3.82	1 937	3.75	51 628	100	
Quintile 3	60 264	74.07	10 336	12.7	6 005	7.38	4 760	5.85	81 365	100	
Quintile 4	59 899	68.04	11 558	13.13	10 347	11.75	6 232	7.08	88 036	100	
Quintile 5 (most deprived)	46 896	60.81	9 199	11.93	13 849	17.96	7 177	9.31	77 121	100	
Heart attack											< 0.001
No	241 071	72.50	37 021	11.13	33 227	9.99	21 199	6.38	332 518	100	
Yes	3 993	83.89	498	10.46	147	3.09	122	2.56	4 760	100	
Stroke											< 0.001
No	238 943	72.39	37 014	11.21	32 965	9.99	21 160	6.41	330 082	100	
Yes	6 121	85.06	505	7.02	409	5.68	161	2.24	7 196	100	
Type 2 diabetes											< 0.001
No	231 038	73.15	33 506	10.61	30 946	9.80	20 369	6.45	315 859	100	
Yes	14 026	65.48	4 013	18.74	2 428	11.34	952	4.44	21 419	100	
Kidney disease											< 0.001
No	231 469	72.26	36 021	11.24	32 040	10.00	20 813	6.50	320 343	100	
Yes	13 595	80.28	1 498	8.85	1 334	7.88	508	3	16 935	100	
Sickle cell disease											< 0.001
No	245 053	72.71	37 515	11.13	33 152	9.84	21 307	6.32	337 027	100	
Yes	11	4.38	4	1.59	222	88.45	14	5.58	251	100	
Thalassemia											< 0.001
No	244 388	72.87	36 860	10.99	33 060	9.86	21 080	6.29	335 388	100	
Yes	676	35.77	659	34.87	314	16.61	241	12.75	1 890	100	
Schizophrenia											< 0.001
No	243 804	72.70	37 346	11.14	32 989	9.84	21 214	6.33	335 353	100	
Yes	1 260	65.45	173	8.99	385	20	107	5.56	1 925	100	
∑	245 064	72.66	37 519	11.12	33 374	9.9	21 321	6.32	337 278	100	

* Note: p-values are obtained from χ^2 tests of independence for each of the variables considered.

Table D.3. Case study 1: unadjusted associations of the response indicator of ethnicity with variables used to inform the imputation of ethnicity; n = 445 199.

Variable	Missing		Observed		∑		p-value
	Frequency	%	Frequency	%	Frequency	%	
Sex							< 0.001
Male	58 598	26.75	160 473	73.25	219 071	100	
Female	49 323	21.81	176 805	78.19	226 128	100	
Age group (years)							< 0.001
0–9	11 418	22.18	40 054	77.82	51 472	100	
10–19	16 960	35.75	30 484	64.25	47 444	100	
20–29	13 868	22.35	48 179	77.65	62 047	100	
30–39	14 975	19.48	61 880	80.52	76 855	100	
40–49	17 821	25.25	52 755	74.75	70 576	100	
50–59	14 285	26.82	38 985	73.18	53 270	100	
60–69	9 243	23.32	30 399	76.68	39 642	100	
70–79	5 360	21.1	20 041	78.9	25 401	100	
80+	3 991	21.58	14 501	78.42	18 492	100	
Townsend score							< 0.001
Quintile 1 (least deprived)	12 642	24.42	39 128	75.58	51 770	100	
Quintile 2	17 015	24.79	51 628	75.21	68 643	100	
Quintile 3	28 180	25.72	81 365	74.28	109 545	100	
Quintile 4	25 381	22.38	88 036	77.62	113 417	100	
Quintile 5 (most deprived)	24 703	24.26	77 121	75.74	101 824	100	
Heart attack							< 0.001
No	106 816	24.31	332 518	75.69	439 334	100	
Yes	1 105	18.84	4 760	81.16	5 865	100	
Stroke							< 0.001
No	106 127	24.33	330 082	75.67	436 209	100	
Yes	1 794	19.96	7 196	80.04	8 990	100	
Type 2 diabetes							< 0.001
No	103 105	24.61	315 859	75.39	418 964	100	
Yes	4 816	18.36	21 419	81.64	26 235	100	
Kidney disease							< 0.001
No	103 856	24.48	320 343	75.52	424 199	100	
Yes	4 065	19.36	16 935	80.64	21 000	100	
Sickle cell disease							0.976
No	107 841	24.24	337 027	75.76	444 868	100	
Yes	80	24.17	251	75.83	331	100	
Thalassemia							0.189
No	107 353	24.25	335 388	75.75	442 741	100	
Yes	568	23.11	1 890	76.89	2 458	100	
Schizophrenia							< 0.001
No	107 486	24.27	335 353	75.73	442 839	100	
Yes	435	18.43	1 925	81.57	2 360	100	
∑	107 921	24.24	337 278	75.76	445 199	100	

* Note: p-values are obtained from χ^2 tests of independence for each of the variables considered.

Table D.4. Case study 2: unadjusted associations of ethnicity with variables used to inform the imputation of ethnicity among the complete records; n = 309 684.

Variable	White		Asian		Black		Mixed/Other		∑		p-value
	Frequency	%	Frequency	%	Frequency	%	Frequency	%	Frequency	%	
Sex											< 0.001
Male	106 727	72.26	17 675	11.97	14 099	9.55	9 207	6.23	147 708	100	
Female	117 676	72.65	17 352	10.71	16 672	10.29	10 276	6.34	161 976	100	
Age group (years)											< 0.001
0–9	21 302	63.28	4 392	13.05	4 522	13.43	3 449	10.25	33 665	100	
10–19	18 769	63.37	3 679	12.42	4 352	14.69	2 820	9.52	29 620	100	
20–29	25 931	68.9	5 078	13.49	3 807	10.12	2 820	7.49	37 636	100	
30–39	37 327	69.28	7 943	14.74	4 926	9.14	3 682	6.83	53 878	100	
40–49	35 829	71.52	5 313	10.61	5 955	11.89	3 000	5.99	50 097	100	
50–59	29 883	75.34	3 748	9.45	4 058	10.23	1 976	4.98	39 665	100	
60–69	25 291	83.42	2 559	8.44	1 453	4.79	1 014	3.34	30 317	100	
70–79	16 771	83.03	1 643	8.13	1 257	6.22	527	2.61	20 198	100	
80+	13 300	91.05	672	4.6	441	3.02	195	1.33	14 608	100	
Townsend score											< 0.001
Quintile 1 (least deprived)	32 770	87.45	2 386	6.37	1 145	3.06	1 173	3.1	37 474	100	
Quintile 2	41 611	84.35	3 897	7.9	1 923	3.9	1 903	3.86	49 334	100	
Quintile 3	56 107	73.57	10 004	13.12	5 678	7.45	4 475	5.87	76 264	100	
Quintile 4	54 545	67.67	10 834	13.44	9 539	11.83	5 682	7.05	80 600	100	
Quintile 5 (most deprived)	39 370	59.64	7 906	11.98	12 486	18.91	6 250	9.47	66 012	100	
Type 2 diabetes											< 0.001
No	212 714	72.96	31 526	10.81	28 660	9.83	18 659	6.40	291 559	100	
Yes	11 689	64.49	3 501	19.32	2 111	11.65	824	4.55	18 125	100	
∑	224 403	72.46	35 027	11.31	30 771	9.94	19 483	6.29	309 684	100	

* Note: p-values are obtained from χ^2 tests of independence for each of the variables considered.

Table D.5. Case study 2: unadjusted associations of the response indicator of ethnicity with variables used to inform the imputation of ethnicity; n = 404 318.

	Missing		Observed		∑		p-value
	Frequency	%	Frequency	%	Frequency	%	
Sex							< 0.001
Male	50 593	25.51	147 708	74.49	198 301	100	
Female	44 041	21.38	161 976	78.62	206 017	100	
Age group (years)							< 0.001
0–9	7 936	19.08	33 665	80.92	41 601	100	
10–19	16 044	35.13	29 620	64.87	45 664	100	
20–29	12 429	24.83	37 636	75.17	50 065	100	
30–39	11 817	17.99	53 878	82.01	65 695	100	
40–49	14 740	22.73	50 097	77.27	64 837	100	
50–59	13 607	25.54	39 665	74.46	53 272	100	
60–69	9 110	23.11	30 317	76.89	39 427	100	
70–79	5 150	20.32	20 198	79.68	25 348	100	
80+	3 801	20.65	14 608	79.35	18 409	100	
Townsend score							< 0.001
Quintile 1 (least deprived)	11 460	23.42	37 474	76.58	48 934	100	
Quintile 2	15 454	23.85	49 334	76.15	64 788	100	
Quintile 3	25 041	24.72	76 264	75.28	101 305	100	
Quintile 4	22 026	21.46	80 600	78.54	102 626	100	
Quintile 5 (most deprived)	20 653	23.83	66 012	76.17	86 665	100	
Type 2 diabetes							< 0.001
No	90 659	23.72	291 559	76.28	382 218	100	
Yes	3 975	17.99	18 125	82.01	22 100	100	
∑	94 634	23.41	309 684	76.59	404 318	100	

* Note: p-values are obtained from χ^2 tests of independence for each of the variables considered.