Multiple Imputation of Missing Data in Multilevel Models

Nidhi Prakash Menon

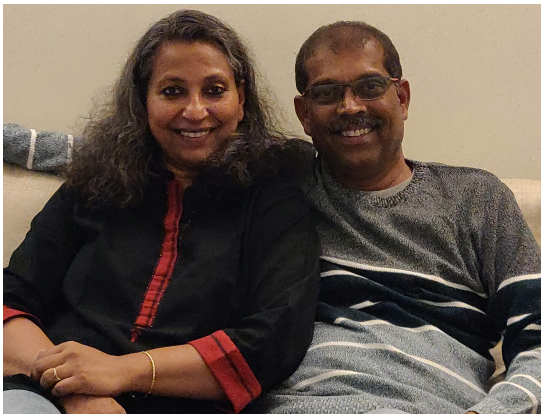
Australian National University

A thesis submitted for the degree of

Doctor of Philosophy

of The Australian National University

March, 2021



This thesis is dedicated to my parents, Prakash and Maya Menon, for their endless love, support and faith in me. They are my strength in everything I do.

Declaration

Candidate

This thesis is an account of research undertaken between January 2017 and March 2020 at The Research School of Population Health, College of Health and Medicine, The Australian National University, Canberra, Australia. I declare that the work contained in this thesis is the result of original research and has not been submitted to any other University or Institution. This thesis contains the following papers to which I made significant contribution as first author:

1. Nidhi Menon, Alice Richardson, Hwan Jin Yoon.Estimating Effect of Maternal and Household Risk Factors Influencing Anaemia among Children under 5 years Using Multiple Imputation. To be submitted for publication.

2. Nidhi Menon, Alice Richardson, Brett Lidbury. The Effect of Multiple Imputation of Routine Pathology Variables on Laboratory diagnosis of Hepatitis C Infection. To be submitted for publication.

3. Nidhi Menon, Binukumar Bhaskarapillai & Alice Richardson (2019) Effect of modeling a multilevel structure on the Indian population to identify the factors influencing HIV infection, *Biostatistics & Epidemiology*, 3:1, 126-139

The details of my contribution to each of these papers is as follows:

1. I surveyed the literature and integrated this with the input and expertise of co-authors to implement our extended multiple imputation model framework to impute missing values of maternal, household and individual variables. I compiled the National Family Health Survey data, and wrote the R code that implemented the model. I contributed to the interpretation of results and produced the figures and tables. I drafted the manuscript and managed the submission process.

2. I developed the study concept. I wrote R code that implemented the model, produced numerical results, and performed imputation and statistical analysis. I contributed to the interpretation of results and produced the figures and tables. I drafted the manuscript and managed the submission process including responses to reviewer feedback.

3. I compiled the National Family Health Survey data, and wrote the STATA code that implemented the model. I contributed to producing and interpretation of the results, made the figures, and drafted the manuscript. I managed the submission process including responses to reviewer feedback.

Nidhi Menon

March, 2021

Acknowledgments

Every PhD journey is unique. My journey has been coloured by a series of wonderful opportunities that I had not fully anticipated when I committed to this journey of three and a half years. The completion of my thesis owes greatly to the support of many, whom I would like to acknowledge here.

First and most of all, I would like to thank my chair and primary supervisor, Associate Professor Alice Richardson for her expertise, guidance and patience throughout this journey. She has patiently endured at least two (sometimes ten!) drafts of almost everything I have written during my PhD. This acknowledgement is probably the only section in this thesis that she hasn't edited. Her encouragement, dedication, attention to every detail and insightful feedback is what led to the completion of my PhD research.

This thesis has benefited greatly from the members of my PhD supervisory panel, Professor Catherine D'Este and Dr Hwan Jin Yoon. I am grateful to Professor Catherine D'Este for her valuable insights, vast experience and constructive feedback which shaped the foundation of my thesis. I also thank her for the coveted opportunity to work with the Australian Commission on Safety and Quality in Health Care on their End of Life Care project. I thank Dr Hwan Jin Yoon, for his continued optimism and help in wrangling through several intangible results. His generosity with his time allocated for our meetings, enthusiasm about my research and constant encouragement are invaluable. I would also like to acknowledge my co-authors, Dr Binukumar Bhasakarapillai and Associate Professor Brett Lidbury for their contribution on the papers listed in this thesis.

I thank the National Computer Infrastructure at the ANU for the opportunity to use their supercomputer, without which many of the analysis performed in the thesis would have been impossible. I also thank NCI Staff Scientist Rika Kobayashi for her time and patience in help-

ing me navigate through the supercomputer.

The life of a student is not without its ups and downs, but was made thoroughly enjoyable with my good friends, including Shivam Mehta, Divyanshu Chauhan and Nikita Ramachandran. I am grateful to Vaishnavi Bharadwaj for her editorial work on this thesis. I thank Madhav Menon for being a wonderful friend and confidante throughout this journey. I would like to acknowledge the community of students and staff at the Research School of Population Health and Graduate House who have been a great source of encouragement and friendship. I thank Graduate House for the opportunity to be a part of their Student Leadership Team which helped me finance my stay on campus. In the last part of my PhD, Associate Professor Bruce Shadbolt employed me to work at the Biological Data Science Institute and the ACT Health Directorate. I am indebted to him for not only providing me with an income but giving me the time and space required to complete my thesis.

My parents, Maya and Prakash Menon, deserve a whole chapter of acknowledgements. They have supported me through every stage of the PhD. They have shared my excitement when things were looking up, encouraged me through disappointments and setbacks, and have celebrated every big and small victory along the way. I thank my brother, Nikhil Menon for stepping in at the time of crisis, and giving me the time and support needed to complete my work. Most importantly I thank my family for helping me keep the right perspective and for being my strength throughout this journey.

Finally, this thesis is submitted at a time of great uncertainty for all, due to the ongoing global battle against the novel coronavirus disease (COVID-19) that is quickly spreading across nations. During these testing times, I would like to express my deepest gratitude and admiration to the healthcare workers, doctors, nurses, technicians, transporters and everyone who is rising to the occasion and supporting patient care.

Abstract

Missing data and its occurrence in statistical analysis has plagued medical and public health researchers for decades. The last few years have seen statisticians and methodologists across the world unite and come together to develop and improve the tools and approaches used in this fight against missing data. This has been made possible due to the increase in implementation of multiple imputation in statistical software, and a stronger push against rudimentary approaches like Complete Case Analysis and single imputation methods. Multiple Imputation (MI) has been in the literature since the 1980s. Despite this, we still observe a huge disparity between awareness and implementation of MI in research. Scientists are aware about the existence of MI, but are hesitant to employ it in practice.

MI has been long recognized as an attractive approach to handle missing values. Despite its early conception and its numerous advantages over the traditional ad hoc methods, there is still limited application of MI in public health research. Multiple imputation methods can be broadly categorized into two: Joint modelling (JoMo) and Multiple Imputation using Chained Equations (MICE). MICE imputes missing data sequentially from a series of univariate distributions while JoMo draws imputations for missing data simultaneously from a multivariate distribution. While these methods have been extensively studied in single level (and some two level contexts), there is limited knowledge about the performance of these methods to handle missing values in more than two levels.

The theory of multiple imputation requires the sampling design be incorporated in the imputation process. Not accounting for complex sample design features, such as stratification and clustering, during imputations can yield biased estimates from a design-based perspective. Most datasets in public health research show some form of natural clustering (individuals within households, households within the same

district, patients within wards, etc.). Cluster effects are often of interest in health research. Missing values can occur at any level in multilevel data, but guidance on multiple imputation in data with more than two levels is currently an open research question.

This thesis implements and extends the Gelman and Hill approach for imputation of missing data at higher levels by including aggregate forms of individual-level measurements to impute for missing values at higher levels. We demonstrate how we can implement this extension in MICE and JoMo to impute for missing values in three level data structures, and produce better estimates than ad hoc approaches, when varying proportions of missing values occur together across all levels in the dataset.

In this thesis, we also demonstrate how the performance of extensions of MI to handle missingness in three level dataset varies with the number of clusters at each level in the dataset. This is verified by using simulation studies as a proof of concept. The analytic work and simulations in this thesis results in several practical recommendations to medical and health researchers.

Finally, this thesis highlights the strengths and limitations of multiple imputation for variables in datasets with more than two levels. The results from the simulation study lead to real world clinical and health examples to further illustrate its application. The illustrative examples demonstrate the impact of accounting for data structure in analysis, constructing rich imputation models for clinical laboratory data, and the effect of MI in developing health profiles using national surveys.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Missing data are considered a challenge in research and analysis because a majority of statistical methods of analysis fail to account for missing values. Researchers are often interested in why and how these missing values arise, but handling them using proper methods is both theoretically and computationally challenging. With improper knowledge or tools to handle these missing values, researchers resort to informal and improvised methods to complete the data, resulting in an increase in bias and unreliability of results. Thus, there is an urgent need to understand the various methods to account for missing values in analysis and how to apply them in survey, medical or public health research. As Paul Allison (2001) wisely stated, *"Analysing data that you do not have is so obviously impossible that it offers endless scope for expert advice on how to do it."*

Why is it important to address this topic?

- Ultimately, any researcher conducting any statistical analysis runs into problems with missing values.
- The presence and magnitude of missing values are often not clearly stated in studies, which can have significant effect on the conclusions drawn from the data and introduce bias in the study. Missing data and methods to handle them continue to remain unreported across studies.
- Default methods including Complete Case Analysis are frequently performed

without mentioning it (Karahalios et al., 2012). Often the lost data can introduce bias in the study as the sample may not be a true representative of the population.

- Principled methods of analysing missing data methods such as missing data imputation, direct likelihood and full-information maximum likelihood have long existed, but broad implementation of these methods continues to remain an obstacle to overcome.

Missing data is a challenge for all studies, however this is especially true for epidemiological studies and clinical research. These study designs are often more complex and technical in nature, with researchers measuring several biological parameters across various levels, at multiple time intervals. These issues are further highlighted in complex datasets (van Buuren, 2018). Traditional methods such as Complete Case Analysis not only remove observations at the lower level units, but presence of missing values in variables captured at higher level may result in deletion of all observations nested within that unit. For example, if we observe missing values for information captured at household level, (e.g. number of members in the household), performing Complete Case Analysis may result in deletion of the information captured for all individuals in that household. Given the complexity of hierarchical data structures and the appropriate modelling procedures, multilevel datasets with missing values require special methods to handle missingness.

1.1.1 Why does this thesis matter?

Imputation as an approach to handle missing data has been around for several decades. In comparison to its counterparts, multiple imputation (MI) has gained popularity (and credibility!) as an attractive approach to handle missing values. Statisticians are now advocating for the use of MI as a principled method for making valid inference under our missing data assumptions. Despite its early conception (Rubin, 1987) and its numerous advantages over the traditional adhoc methods (see section 2.2), there is still limited application of MI in public health research.

Karahalios et al. (2012) conducted a systematic review on reporting and handling of missing data in cohort studies. They reviewed 82 cohort studies, of which 66 (80%) reported an overall presence of missing values, 13 (16%) reported reasons for missingness and 14 (16%) did not mention the statistical methodology implemented to handle these missing values. The authors reported 38 of these cohort studies excluded participants with missing data at any repeated waves of exposure, and 54 (66%) of the studies used the Complete Case Analysis method.

This review throws light on the current practices on handling missing data across public health research. While the review focused on cohort studies, we can expect to see similar trends across other study designs. It is positive to note that at least 5 (6%) of the 82 studies reported using multiple imputation methods. This number is very small in comparison to the 54 (66%) studies that used Complete Case Analysis.

Multiple imputation has been in the literature since the 1980s (Rubin, 1987). It has recently gained popularity in the statistical community with its implementation across various statistical tools. Despite these developments, a barrier continues to exist when it comes to implementing MI in practice. This could more likely be due to researchers having a limited understanding of the theory of MI, leading to lacking confidence to use it in practice. This thesis aims to make MI more accessible to researchers by providing a breakdown of how one can practically implement MI for two and three level datasets using R (see Chapter 4). It also provides sufficient evidence and illustrations to convince public health researchers to utilise their data to the maximum by steering away from Complete Case Analysis and embracing the use of MI (see Chapter 5).

The theory of multiple imputation requires that imputations be subject to the sampling design employed in the collection of data. Not accounting for complex sample design features, such as stratification and clustering, during imputations can yield biased estimates from a design-based perspective. In this thesis, we dive into how to incorporate hierarchical study designs in both the imputation and analysis model,

specifically focusing on three level datasets. As a result, we also discuss methods of handling derived variables, such as cluster means and deviations in imputations. We provide an illustration of the proposed extension of the method, to impute for missing data in national surveys where children are clustered within mothers, who are further clustered within households. Information is captured (and not captured) at each level in this dataset. Thus the structure of the dataset is important not only for analysis (as illustrated by Case Study 2 in Section 7.3), but also when developing the imputation models.

Most datasets in public health research show some form of natural clustering (individuals within households, households within districts, patients within wards, etc.). There is a general consensus among researchers that this clustering of data should be accommodated in the analysis, to account for within cluster variability and other cluster effects. Researchers are often interested in the relationship between the variables observed at these higher levels and the outcome variable. For example, researchers maybe interested on the impact of biological, behavioural, community and environmental risk factors on a person's physical and emotional well-being. To accomplish this, multilevel models are commonly applied in practice, where household, community and environmental risk factors are modelled as random effects. However, despite the abundant use of multilevel models in health and medical research, the implications of handling non-response in these models have never been thoroughly studied. This problem worsens in the Complete Case Analysis of hierarchically structured models, where missing values at higher level units could result in exclusion of all observations nested within that unit.

Three level models frequently occur across medical and health research where information captured for individuals are nested within households that are further nested within districts, or for patients nested within doctors within wards, or individuals nested within households nested within districts. In each of these scenarios, researchers run the risk of information not being captured at each of these levels. The repercussions of an adhoc approach to handle missing data in these scenarios

are severe. For example, missing values for district level variables could result in the exclusion of all households that were captured for that particular district; or dropping of relevant variables from analysis due to high proportions of missingness. Despite MI being widely accepted as the preferred method to handle non-response, research on the imputation strategies especially in the context of three level data structures is limited.

1.2 What does this thesis entail?

The thesis presents a two-fold understanding of extending MI in multilevel models. First, we offer simulation results discussing the performance of MI after extending MICE and JoMo to handle variables captured up to a third level. Second, we offer an illustrative guide on the application of MI for users who are familiar with the theory of MI but are hesitant to apply it in their analysis. The central chapters of this thesis are ordered such that every chapter extends the ideas presented in the previous chapters. Three core aspects are examined in this thesis.

1. The current theory of multilevel multiple imputation was developed by Schafer and Graham (2002), Goldstein, Pan, and Bynner (2004), Carpenter and Kenward (2012), Quartagno and Carpenter (2016) and Audigier et al. (2018). The theory was incorporated by Yucel in (2008), and Gelman and Hill in (2006) which was restricted to two level hierarchical models. Three level models frequently occur across medical and health research where information is captured for individuals are nested within households that are further nested within districts, or for patients nested within doctors within wards, or children nested within mothers nested within households. In each of these scenarios researchers run the risk of information not being captured at each of these levels. There is an urgent need to extend the current theory to incorporate three level data structures.

2. There are limited resources that demonstrate the practical implementation of popular methods of multiple imputation such as MICE (van Buuren et al., 2015)

and JoMo (Quartagno & Carpenter, 2016) in the multilevel context. In addition, the current tools are limited to imputation for two level datasets only with the exception of BLIMP software developed by Enders, Keller, and Levy (2018). Existing imputation routines are diverse and offer different functionality; all approaches can accommodate basic random intercept analyses but they differ in their functionality to estimate random slopes,binary and ordinal variables, within cluster and between cluster covariance matrices, and partially observed covariates at higher levels (Enders et al., 2018). Recently, there have been calls to extend the current software to improve the imputation procedures made available to researchers working with three-level datasets. BLIMP currently remains the only tool that successfully imputes missing values in the third level. However, the software remains untested for imputation of derived variables in high level units. There is a need to extend the current implementation of MI and test the performance of this extension.

3. The evidence for the impact of missing data occurring together across all levels in the dataset is limited with a majority of studies focusing on missing values occurring in one level only. There has been no study on the performance of imputation procedures for derived variables such as cluster means and deviations for multilevel models.

Finally, this thesis demonstrates real world application of multiple imputation in three level clustered datasets. This thesis is among the first few to study the impact of multiple imputation when missing values occur together across all levels in the data.

1.2.1 What does this thesis not entail?

It is important to specify at the outset what the boundaries of the thesis will be. Hierarchical structures also include repeated observations taken on an individual subject. The observations taken on the same individual form the various levels of hierarchy and the first observation forms the first level in the multilevel analysis. This thesis does not focus on longitudinal or repeated measures datasets. The structure of the

datasets are solely hierarchical clustered.

Likewise, this thesis focuses exclusively on hierarchical simulated datasets, where each observation is nested within one and only one higher level unit. Sometimes, health outcomes or medical situations do not have a hierarchical structure. The results of the thesis cannot be extrapolated to non-hierarchical multilevel modelling. These include *Multiple Membership* and *Cross Classification* models. In Multiple Membership models, a lower unit is allowed to be a member of more than one higher level unit; for example, patients staying in more than one nursing unit during their hospital stay. In Cross Classification models, lower level units are cross-classified into two or more higher levels. For example, patients are nested in nursing units but at the same time, the patients in that nursing unit are affiliated to different doctors. But, nursing units are not nested in doctors and doctors are not nested in nursing units; the two clusters exist at the same level (Greenland, 2000). The application of the methods implemented in this thesis have not been tested for such data structures and it would be unwise to extrapolate the results observed in this thesis to a non-hierarchical model. More research is required to check the validity of these methods on Multiple Membership and Cross Classification models.

1.3 Thesis Structure

This thesis consists of eight chapters including the Introduction. Chapter 2 lays out the background information necessary to understand multiple imputation and the various characteristics of missing data. This chapter discusses rudimentary and existing methods of handling missing data across literature. The concept of multiple imputation and the methods adopted in this thesis are also introduced in the chapter. Chapter 3 presents the existing literature on the intersection between missing data and multilevel models. The gaps in the current literature are identified in this chapter. The chapter also defines the goals of the thesis.

Chapter 4 is at the core of this thesis. In this chapter, we extend the current prac-

tice of MICE and JoMo to incorporate the three level data structure. This chapter attempts to fill in the gap in literature on the practical implementation of multilevel multiple imputation using statistical software, and provides an illustrated guide of the algorithm implemented. The structure of the simulation study is also explained here. This structure is the basis of the analysis in Chapters 5 and 6, and provides the basic framework of the imputation models implemented across the thesis.

Chapter 5 provides a verification of the performance of the extended methods of MI, using the simulation framework described in Chapter 4. This chapter also zooms into some of the gaps that became visible through the results obtained from the simulation study, and provides key insights into the limitations and strengths of the extended methods. Not only does this chapter provide a great insight into the performance of the extended MI methods; referred within this thesis by the suffix (ext.); but also provides an understanding of the approaches to be used to overcome some hurdles we might encounter when imputing multilevel data.

Chapter 6 addresses a gap in the literature on the impact of varying cluster sizes and number of clusters on MI methods. This chapter particularly comments on the robustness of multilevel multiple imputation when used to handle missing data in varying cluster sizes. This is among the first few studies to look into varying cluster sizes across three level data structures to study the impact on MI. In addition, the chapter also provides an adjustment bias to the interval estimates obtained from MI (ext.).

Chapter 7 consists of three case studies illustrating the impact of using multilevel models and multiple imputation in analysis, in real world datasets. These case studies are manuscripts that have been published or submitted for publication in peer reviewed journals. The status of each paper and my personal contribution towards developing the papers has already been listed in the Declaration.

The first case study is an illustrative example lying at the intersection of both multiple

imputation and multilevel modelling and applies the imputation models described in Chapter 4 to profile maternal and child anaemia levels in national datasets.The second case study draws out the impact of multilevel modelling for three level data structures as described in Chapter 3. The final case study demonstrates the use of multiple imputation methods in handling missing values in laboratory data. Each case study is prefixed by an Introductory section describing the paper. The thesis concludes with Chapter 8 which discusses the key findings in the thesis, provides recommendations for applied researchers, states limitations of the methodology and discusses directions for future research.

Chapter 2

Missing Data

This chapter describes the importance of addressing missing values in a study, the different types of missingness, patterns of missingness, and rudimentary and modern methods of handling missing data including both single and multiple imputation. It is important that the reader understands these concepts as they are discussed in more detail in further chapters. In this chapter, we will be providing an overview of the origin of multiple imputation; the different methods of multiple imputation; and the implementation of these techniques in complex data structures, specifically multilevelled clustered data.

Sections 2.1 and 2.2 provide an overview of underlying concepts and theories, and common practices to handle missing data. They focus on the different dimensions of missing data, various single imputation practices and their limitations. It is important to understand the different dimensions of missing values we need to consider before proceeding into multiple imputation. Despite its various limitations, single imputation has provided researchers with the necessary stepping stones to develop the idea of multiple imputation. As demonstrated by Karahalios et al.(2012), even with their many limitations, single imputation methods continue to remain a popular approach among researchers when handling missing data.

Section 2.3 provides an in-depth understanding of both the concept of multiple imputation, as well as an overview of popular methods of multiple imputation. In addition, sections 2.3.8 and 2.3.9 addresses key concerns one should consider before proceeding with multiple imputation. These sections highlights topics such as

imputation of derived variables and uncongeniality of imputation models, that are frequently overlooked by imputers. Finally section 2.4 provides a summary of concepts covered in this chapter and a short insight into what to expect in the next chapters.

2.1 Principles of Dealing with Missing Data: Background

Researchers and methodologists frequently encounter the problem of missingness. Missing values threaten the representativeness of the sample and restricts the researcher's ability to infer on the target population (Groves, 2004). There are six main factors to consider when dealing with missing data: type of missingness, pattern of missingness, mechanisms of missingness, study design, location of missingness, and the distribution of the variable. Before we look into the above factors, it is important to define what we mean by missingness.

2.1.1 When do we say a value is missing?

Data is said to have missing values when no value is stored for one or more variables in an observation. This could be due to a variety to reasons. Information could be missing due to multiple reasons. In cross sectional studies, missing values could be because the individual failed/chose not to answer certain questions, or it could be due inaccessibility to certain remote locations. Some missing values could be due to the construction of the questionnaire. For instance, if a certain item on the questionnaire is addressed to only a subset of the sample. For example, if the question asked is "Do you smoke? If yes, proceed to the next question, else skip to the next section". In such cases, missing values for non-smokers will be observed for some variables. Questionnaires often contain codes to indicate responses such as "Not Applicable", "Don't Know", "Unsure", etc. If the study contains measurements drawn at regular intervals, it is possible that missing values occurred due to a change of location, other personal commitments, and treatment emergent adverse events and on rare occasions due to death.

Due to the numerous causes for missingness, it is important to understand why

missing values occurred in the first place before we proceed with analysis and management of missing data. Studies often and rightly cite the reasons for the occurrence of missing values, giving us an insight into not only the mechanisms of missingness, but also a gap or a kink in the processes in place for data collection.

2.1.2 Types of missingness

Missing values often present themselves as non-response in surveys. Non-response can be either *Item* or *Unit* non-response.

Rubin (1987) explained *Item non-response* to be when individuals in the study choose to not respond to certain questions asked. *Unit non-response* is when the subject chooses not to respond to any of the questions in the study, or the individual refuses to participate in the study. Item non-response increases further risk to inference sometimes compounding on unit non-response. Each type of non-response presents itself differently in the data. Item non-response would result in missing values across variables (conventionally captured as columns).

Unit non-response would result in missing values across individuals (conventionally captured as rows) in the dataset. Surveys often use weights to adjust for non-response in the analysis. Unit non-response would result in a direct reduction in sample size. In this thesis, we limit the discussion to item non-response. Non-response introduces bias in estimates, especially when there is a difference in characteristics measured between the respondents and the non-respondents. Further, it increases the total variance of the estimates as the sample size is reduced from what was originally sought.

This thesis is limited to handling item non-response. This decision was made as item non-responses are a frequently occurring phenomenon in large cross sectional sur-

veys. Several demographic surveys often use weights to adjust for unit non-response in the analysis, however limited efforts are made to adjust for item non-response. This issue is further aggravated when item non-response occurs across multiple levels in hierarchical data structures. This thesis highlights measures and methods to address item non-response in the context of multilevel models. Unit non-response currently sits beyond the scope of this thesis.

Missing values also frequently occur in longitudinal studies. Here, they present themselves differently. In longitudinal or prospective studies, we observe that information for individuals are missing at certain time-points and available otherwise. This type of missingness is referred to as *Wave non-response*. Loss to follow-up is a special type of wave response where the individuals drops out of the study. Loss to follow-up can severely compromise the validity of a study (Dettori, 2011). Multilevel structures can be either longitudinal or hierarchical in nature; or data can be both multilevel (hierarchical) and longitudinal, as the repeated measures on individuals creates a hierarchy. While it is important to note that missing data presents differently in different study designs, it is important to note that this thesis focuses on the latter and thus does not focus on the issues surrounding loss to follow-up. This discussion can be found in other literature sources (Lee et al., 2016; De Silva et al., 2019; Cro et al., 2016).

2.1.3 Mechanisms of Missingness

Rubin (1976) classified missing values into 3 types- *missing completely at random (MCAR)*, *missing at random (MAR)*, and *missing not at random (MNAR)*.

Data is said to be missing completely at random (MCAR) if the probability of a value being missing is unrelated to the observed and missing data for that unit. This means that regardless of the observed values, the probability of missing is the same for all units. Administrative variables that are routinely collected in surveys, play a key role in understanding the mechanisms of missingness. These variable often serve

as auxiliary variables that can be predictive of values being missing, but not predictive of actual missing values. In such cases, we can say that the data are MCAR with respect to the scientific question, even though we have variables that are predictive of values being missing.

Unlike standard notation, in this thesis we refer to variables with the subscripts i,j and k to signal to the level at which the variable is captured. For example, X_{ijk} is a variable captured at the lowest level (Level 1) where (i) refers to the number of units at the lowest level, (j) refers to the second level (Level 2) and (k) refers to the third level (Level 3). This is done to help readers navigate through the multiple levels referenced through out this thesis.

Data is said to be missing at random (MAR) if the probability of missingness depends on the observed values but not the missing values. This implies that probability of missingness in a variable, say X^{mis}, may be dependent on an observed variable, say X^{obs}. This process can be modelled as a logistic regression when response in the outcome variable takes values '1' if observed and '0' if missing (see section 4.2.1 to know how we generated MAR in this thesis). It is important to note that under the MAR assumption, the chance of the data being missing depends on the missing value, but this association can be broken using the observed data, hence it is a conditional independence statement. Finally, if the MAR assumption is violated, then the data is said to be missing not at random (MNAR).

Mathematically, this can be explained using the standard notation as described by Rubin (1987). Consider a dataset where X^{obs} is the observed part of the dataset and X^{mis} is the missing part. We define a response matrix R where $R = 1$ when the variable is observed and $R = 0$ when the variable is missing. Here, R is a matrix of 1's and 0's, and stores the location of missing data in X.

So, we define the situation MCAR as $P(R|X) = P(R)$. The probability is not dependent on X^{obs} or X^{mis}. Similarly MAR is defined as $P(R|X) = P(R|X^{obs})$. As defined

above, we see that the probability of missingness depends on the observed values in X when data is MAR. The situation MNAR is defined as $P(R|X) \neq P(R|X^{obs})$. In this thesis, we restrict ourselves to the MCAR and MAR mechanisms only (see section 5.3).

It is advised that researchers use both statistical methods and contextual knowledge to draw conclusions about missing-data mechanisms. Observed data can be often used to inspect the credibility of different missingness assumptions. However, it is critical that statistical analyses is underpinned by contextual knowledge to judge the overall plausibility of a missingness assumption. If the assumed missingness mechanism satisfies the MAR assumption, then MI can be used to obtain more efficient estimates (Bartlett, Harel, & Carpenter, 2015).

2.1.4 Ignorability in Missing Data

Ignorability of missing data mechanisms can be illustrated by looking at the role of the response matrix R in the model used to impute the variable with missing values, commonly referred to as the *imputation model*. Using the notations explained above, R =1 implies that the variable is present and R_{ij}=0 implies that the variable is missing.

van Buuren (2007) explains the concept of ignorability further. The conditional distribution $P(X^{mis}|X^{obs}, R)$; summarizes the information about X^{mis} that is present in X^{obs} and R. When the response mechanism is ignorable, we can equate $P(X^{mis}|X^{obs}, R = 1) = P(X^{mis}|X^{obs}, R = 0)$. $R = 0$ implies that the cases have missing values in X and thus cannot provide any information on $P(X^{mis}|X^{obs}, R)$. However, $R = 0$ also implies that there is a possibility that $P(X^{mis}|X^{obs}, R = 1) \neq P(X^{mis}|X^{obs}, R = 0)$ (Rubin, 1987, p. 205).

As stated in his book van Buuren (2018); *'There is nothing in the theory of MI that*

prevents appropriate inferences under $P(X|X^{obs}, R = 0)$. *MI will also work for non ignorable response mechanisms. The assumption of ignorability is critical in that we assume the available data is sufficient to correct the effects of missing data"*.

If the mechanism is ignorable i.e. $P(X|X^{obs}, R = 1) = P(X|X^{obs}, R = 0)$ then this suggests that the distribution of X for the response and non-response groups is the same. This means that we can use the observed data to construct the posterior distribution for X^{mis} i.e. $P(X|X^{obs}, R = 1)$ and generate imputations for the missing values in X. Conversely, if the mechanism is non-ignorable, we should include R in the imputation model. The assumption of ignorability cannot be tested on the data by itself. Ignorability also has a further assumption of parameter distinctness (i.e. the parameter governing the complete data distribution is distinct from that governing the missing data distribution (Rubin, 1976). Researchers should bear in mind that the construction of non-ignorable models should be based on the mechanisms that created the missing values in the data.

2.1.5 Patterns of Missingness

Patterns of missingness can be monotone, intermittent or mixed. These are commonly applicable to longitudinal studies. A monotone pattern is observed when a patient drops out of a study. Hence data is available for all assessments till he/she drops out. An intermittent pattern is observed when missing values are present between observed assessments. (Fielding, Fayers, & Ramsay, 2009).

It is important to distinguish between patterns, locations and mechanisms of missingness. This thesis focuses on the various mechanisms of missingness across different locations in the dataset.

2.1.6 Study Design and Location of Missingness

The type of study design would define the structure of the dataset. For example, a multi-centre study or any longitudinal study design would have a clustered data

structure. In comparison, a cross sectional survey would have a flat data structure. This in turn, would impact where the variable with missing values is located. For example, if missing values are observed in a variable captured at a higher level this would impact how we define the variables that would be used as predictors to impute the missing variable. In addition, the structure of the imputation model should reflect the data structure and the analysis model. Thus study design, and location of the missing values are key factors to bear in mind prior to imputation. This is discussed further in section 2.3.5.

2.1.7 Distribution of Missing Variable

The general guidelines that one follows when choosing the correct analysis model is adhered to when choosing the right model to impute for missing values. For example, a logistic regression model is used when imputing for missing values in binary variable, a poisson model is used to impute for missing values in count data, and a linear regression model is used to impute for missing values in continuous variables. There are key points that we bear in mind when identifying the covariates to be added in to these models. This is discussed in Section 2.3.7 and an algorithm (see Table 4.1.1) is provided to help further breakdown the steps involved.

2.2 Strategies for Dealing With Missing Values

2.2.1 Prevention

"Obviously the best way to treat missing data is to not have them" – (Woodbury & Orchard, 1970, p. 697).

Missing data experts van Buuren (2018) and McKnight, McKnight, Sidani, and Figueredo (2007) have also concurred with Orchard and Woodbury when they rightly state that the best solution to handle missing data problems is not have any missing values. However, it is important to recognize the impracticability in trying to achieve the same. Even the most carefully designed and executed studies produce missing val-

ues. It is better to reduce the amount of missing values in the dataset by dealing with any unintentional missing data. There are several factors influencing the occurrence of missing values in the study including design of the study, method of data collection employed, indisposed participants, untraceable participants, non-response, etc. Better planning and execution of any research after accounting for these potential factors would help substantially reduce missing values in the study.

2.2.2 Complete Case Analysis

The most commonly used method of dealing with missing observations is to use only those cases with complete information. This is employed by most statistical software as they filter all observations with missing values. This method is commonly referred to as Complete Case Analysis.

The first issue that we observe with Complete Case Analysis is whether the complete cases can be considered a random sample of the whole sample. If the data is missing completely at random (MCAR), then Complete Case Analysis can be considered as a random sample of the sample selected, and provides unbiased estimates of mean and variance. However, since data that is MCAR is a rare occurrence in real-world data, it is almost always the case that cases with complete data for the variables included in the analysis are not representative of the whole sample. This means that subjects who have missing information are different from the subjects with complete information. This difference can lead to bias in estimation of various parameters, when Complete Case Analysis is employed (Nguyen, Carlin, & Lee, 2017a).

As above, Complete Case Analysis is valid under a more general mechanism than MCAR, although it will frequently be inefficient – unless only outcomes are incomplete. While marginal statistics are always biased if data are not MCAR, regression analyses will not be, if the probability of a complete record is driven by the independent variables, and given these, does not involve the dependent variable in the regression. A more detailed discussion on when to use MI and when Complete Case

Analysis is enough, is provided in section 2.3.1. Another, issue that arises due to Complete Case Analysis is the loss of power due to the reduction in the sample size. This is more obvious when the proportion of missing observations is large or if number of variables is large. This does lead to the larger question about whether the resistance towards implementation of imputation techniques (both single and multiple), compared to Complete Case Analysis is largely due to its default setting across statistical software.

2.2.3 Single Imputation

An alternate approach to removing any or all observation with missing observations is to fill in or "impute" the missing values. Several imputation approaches have been developed over the years, some simple, others more complex. These methods retain the whole sample which can improve both bias and precision, in comparison to Complete Case Analysis. Commonly employed methods of single imputation include *'Mean Imputation'*, *'Last Observation Carried Forward (LOCF)'* (Gelman & Hill, 2006; Streiner, 2008),*'Regression Imputation'* and *'Hot Deck Imputation'* (Ono & Miller, 1969).

The different types of Single Imputation methods mentioned above are explained further.

1. **Mean Imputation:** This is possibly the easiest method of imputation. It involves replacing a missing value by the mean value of the variable. This imputation however can distort the distribution of the variable if the proportion of missing values is high and the variable contains extreme observations. The mean is highly influenced by outliers and extreme values. This could result in complications with estimated summary measures. Mean imputation shrinks standard errors thus invalidating most hypothesis tests, and also distort relationships between variables such as correlations. This affects standard errors, confidence intervals, and other inferential statistics.

2. **Last Observation Carried Forward:** This method of imputation can be used in

longitudinal studies. The last observed non-missing value is used to fill in for missing observations. The researcher needs to bear in mind that missing data due to drop-outs in any clinical study can be classified as non-informative (this is essentially random where the subject changes location) or informative (not random; where the subject drops out due to adverse effects which could be treatment emergent) (Gelman & Hill, 2006). Because one can never be certain whether data is non-informatively or informatively missing, it is considered good practice not to ignore drop-outs. Last observation carried forward (LOCF) is a commonly used method of imputing data with drop-outs. LOCF uses the last value observed before drop-out, regardless of when it occurred.This method of imputation assumes that the response remains constant at the last observed value.

If a person drops out of a longitudinal study, their last valid data point replaces all consecutive missing observations. This method assumes there will be no further change in the person's measurements, thus underestimating or over estimating what the person's actual value would have been had they continued to participate in the study.
Because it is often the case that more patients drop out of the placebo condition than the control arm, LOCF may actually result in a bias in favour of the alternative hypothesis (Streiner, 2008).

3. **Regression Imputation:** Regression imputation incorporates the information from other variables in order to impute for missing values in the target variable, by fitting a regression model. Predictions for missing values in the outcome variable are then calculated using the model fitted. Regression imputation produces unbiased estimates of the mean when data is MAR and the independent variables in the model are non-missing. *The regression estimates are unbiased under MAR if the factors that influence the missingness are part of the regression model* (van Buuren, 2018). This is important as we use this theory to influence our construction of the imputation model in MI. This ties back to the concept of de-

veloping a flexible and congenial imputation model that accounts for variables driving missingness in partially observed covariates (see sections 2.3.8, 2.3.7).

4. **Hot Deck Imputation:** Hot deck imputation is a method for handling missing data in which each missing value is replaced with an observed response from a 'similar' unit. It involves replacing missing values of one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor) that is similar to the non-respondent with respect to characteristics observed in both respondents. The hot deck method is often preferred as it does not rely on model fitting for the value to be imputed, and thus is potentially less sensitive to model misspecification. However, the method makes inherent assumptions through the choice of metric to match donors to recipients, and the variables included in this metric, so it is not assumption free (Andridge, 2011).

There are two major attractions supporting use of single imputation. First, standard complete data methods of analysis can be applied on the imputed dataset and second, the efforts in constructing the single imputations for the data need to be carried out by the imputer only once. While LOCF and Hot Deck Imputation are still widely practised in clinical research, single imputation methods still fail to capture variability in the data due to missing values and reduce efficiency of estimation (Rubin, 1987). However one of the biggest limitations of single imputation is that we replace a missing value by a single value thus treating it as if it were known with certainty. Another way of saying this is that the statistical analysis software cannot distinguish between the observed and imputed data - but they have very different status, because the imputed data is just a single draw from the estimated distribution of the missing data given the observed, and thus has much greater variability. As a result, single imputation disregards any uncertainty in the available complete data and consequently underestimates the variability. This limitation is naturally overcome by implementing multiple imputation which is discussed in the next section.

2.3 Multiple Imputation

Multiple imputation (MI) is now being recognised as the recommended method to impute for missing values. It was first proposed by Rubin in 1987 in his book *'Multiple Imputation for Non response in Surveys'* and was recommended as a tool statistical agencies could use to handle non-response in large datasets, and disseminate results to the public. Each missing value is replaced by two or more imputed values in order to represent the uncertainty about which value to impute (Rubin, 1987). Standard statistical methods are then applied to each of these imputed datasets and the results are pooled from these analysis. MI therefore consists of 3 steps namely imputation, analysis and pooling (see Fig 2.3.1).

Imputation: The missing values in the incomplete dataset are imputed m times. The default value of m is 5. However, the researcher can change this value as needed. This is discussed in greater detail in section 2.3.5. The imputed values are drawn from a distribution. This step results in m complete (imputed) datasets.

Analysis: Standard statistical tests are performed on each of these m imputed datasets. This would result in m analyses and m sets of estimates from the analyses.

Pooling: Final inference on the regression coefficient estimates is made by averaging the estimates obtained from each of the m imputed datasets. This is done as shown by Little and Rubin (2014) and reproduced in the next paragraph.

Let $\hat{\beta_m}$ be the multiple imputation estimator of β; and s_m^2 be the standard error for the β estimate obtained from the analysis performed on each of the M datasets, where $m = 1, 2, ..., M$. i.e. $\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta_m}$

The final MI variance estimate is given as $\hat{W} + \hat{B}$; where,

$$\hat{V}_\beta = \hat{W} = \frac{1}{M} \sum_{m=1}^{M} s_m^2$$

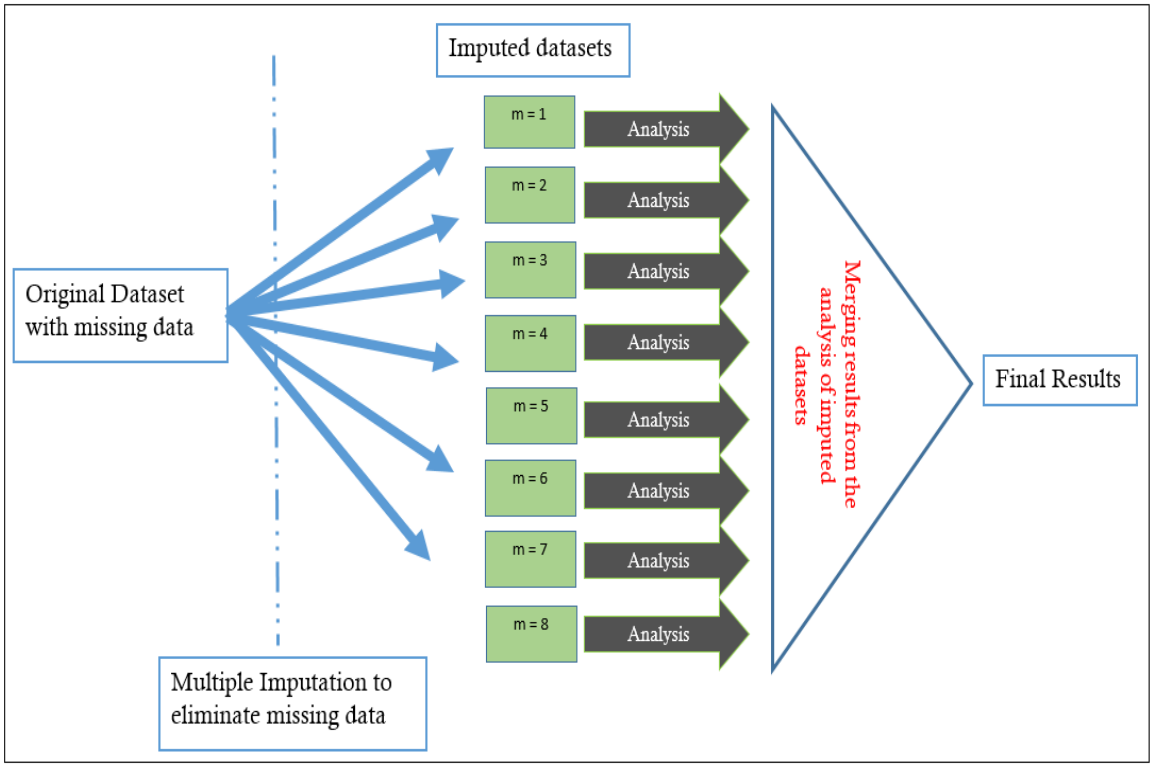


Figure 2.3.1: Diagrammatic Representation of the MI process (example: m = 8)

and

$$\hat{B} = \frac{1}{M}\sum_{m=1}^{M}(\hat{\beta_m} - \hat{\beta})^2$$

2.3.1 When do we use Multiple Imputation?

As mentioned previously, statisticians are now advocating for the use of MI as a principled method for making valid inference under when missing values are observed in the dataset. Does that mean that we *always* need to use multiple imputation when there are missing values? In his book *"Flexible Imputation of Missing Data"*, van Buuren (2018) has recommended some scenarios where alternate methods to multiple imputations are preferred. Complete Case Analysis can be a good substitute to multiple imputation in certain situations. If the mechanism is MCAR, then a Complete Case Analysis provides unbiased estimates (see Table 5.4.3). In addition, if the complete data model is a regression model with missing values only

in the outcome variable, then Complete Case Analysis and multiple imputation are equivalent. However, Complete Case Analysis reduces its efficiency if missing values occur in the independent variables. Complete Case Analysis outperforms multiple imputation under two special scenarios (van Buuren, 2018, p. 198). The first scenario is if the probability of missingness does not depend on the outcome variable. When this assumption is satisfied, the regression coefficients obtained from the Complete Case Analysis are unbiased. This holds true across any type of regression model and for missing data in both independent and outcome variables. It is important to acknowledge that even if a Complete Case Analysis is unbiased, MI can still improve efficiency hence may still be preferable over a Complete Case Analysis

The second scenario specifically applies to a Complete Case Analysis performed using a logistic regression model (Bartlett, Harel, & Carpenter, 2015). Consider a scenario where missing values occur either in a dichotomous outcome or predictor variable (not both!). The regression estimates (except the intercept) generated using Complete Case Analysis are unbiased only if the probability of missingness depends on the outcome variable and not on the independent variable. ***This property fails to hold true if missing values are present in both the outcome and predictor variables!***

Despite these properties of biasedness and unbiasedness, there still appears to be a gap in the knowledge of use of multiple imputation by clinicians and health researchers. Many studies continue to remove any observations with missing values with little to no attempt to try and treat for missingness. The consequences of this could often be the reduction in statistical power due to a reduction in sample size, and biased results.

Before analysing data with missing values, researchers should consider the mechanism by which the data has gone missing (see section 2.1.3) and perform analysis in alignment with these assumptions. Before we proceed with identifying predictors

for the imputation model, it is important to define what we mean by an imputation model.

Imputation Model & Analysis Model: An imputation model is a model that is developed to impute for missing values in the dataset. This is not to be confused with the analysis model which is developed in order to answer the research question asked. The imputation model should reflect the sampling design and strategy employed in the study. As Meng (1994) wisely said, *"Sensible imputation models should not only use all available information to increase predictive power, but should also be as general and objective as practical in order to accommodate a potentially large number of different data analyses."*

In practice, multiple imputation can be implemented in several ways. Popular implementations of multiple imputation in software, include Multiple Imputation using Chained Equations (MICE) and Joint Modelling (JoMo). MICE imputes variables with missing values one at a time from a series of univariate conditional distribution. In contrast, JoMo draws missing values simultaneously for all incomplete variables using a multivariate distribution. Both methods are explained in detail below.

2.3.2 Multiple Imputation using Chained Equations (MICE)

Multiple imputation using chained equations (MICE) is a powerful method for generating imputations, especially in large data sets with complex structures, consisting of both categorical and continuous variables. It defines a multivariate imputation model on a variable-by-variable basis. MICE does not assume the joint distribution of the data is multivariate normal. The variables are reordered to obtain a pattern of missingness as close to monotonous as possible.This allows the user to maximise the information available in the dataset that can be used for imputation.

2.3.3 How does MICE work?

Using the notations from van Buuren (2007), suppose $X = (X_1, X_2 \dots X_p)$ is a vector of p random variables with p-variate distribution $P(X|\theta)$. We assume that the joint multivariate distribution of X is completely specified by θ, a vector of unknown parameters. Suppose, X is multivariate normally distributed, $\theta \sim (\mu, \Sigma)$, with μ (a p-dimensional mean vector) and Σ (a p x p covariance matrix). Let the matrix $x = (x_1 \dots x_n)$ with $x_{ij} = (x_{i1}, x_{i2} \dots x_{ip}), i = (1...n)$ and $(j = 1, 2, ...p)$ be a sample of the matrix X .

Each column in X has missing data. The standard procedure for creating multiple imputations x^* of the matrix X^{mis} is as follows:

1. Calculate the posterior distribution $P(\theta|X^{obs})$ of θ based on the observed data X^{obs};

2. Draw a value θ^* from $P(\theta|X^{obs})$;

3. Draw a value x^* from $P(X^{mis}|X^{obs}, \theta = \theta^*)$, the conditional posterior distribution of X^{mis} given $\theta = \theta^*$. The X^{obs} that are conditioned on becomes the imputed values.

Running through steps 2) and 3) for all variables is referred to as a cycle. At the end of the first cycle, all missing values have been replaced by imputed values. A number of cycles are run to let the algorithm converge.

To further explain this process, Keogh, Seaman, Bartlett, and Wood (2018) focused on one of the most frequently observed scenarios with missing observations in some of the covariates. They considered a scenario where $X = (X_1, ..., X_p)$ is a set of partially observed covariates and X^{obs} contains all completely observed covariates. The outcome variable Y has no missing values. In this case, every partially observed covariate X_j has an imputation model $f(X_j|X_{-j}, Z, Y; \theta_j)$. To elaborate this further, we consider a scenario with 3 partially missing covariates namely X_1, X_2 and X_3 where $X_1 = [X_1^{mis}, X_1^{obs}]$, $X_2 = [X_2^{mis}, X_2^{obs}]$,& $X_3 = [X_3^{mis}, X_3^{obs}]$, and θ_1, θ_2 and θ_3 are the

parameters to be estimated.
Thus the imputation algorithms will be as follows:

Iteration (1):

$$\theta_1^{(1)} \sim f(\theta_1).f(x_1^{obs}|x_2^{obs}.\, x_3^{obs}.\, y.\, \theta_1)$$

$$x_1^{mis(1)} \sim f(x_1^{mis}|\, x_2^{obs}.\, x_3^{obs}.\, y.\, \theta_1^{(1)})$$

The underlying imputation model for $x_1^{mis(1)}$ is as follows:

$$x_1^{mis(1)} = \theta_{10}^{(1)} + \theta_{11}^{(1)} x_2^{obs} + \theta_{12}^{(1)} x_3^{obs} + \theta_{13}^{(1)} y + e_1$$

In the first step, we define a posterior distribution for θ_1 based on the observed data in x_2^{obs} and x_3^{obs}. Parameter θ_1 corresponds to the first variable X_1. The value of θ_1^* is drawn from this posterior distribution. We then draw a value for $x_1^{mis(1)}$ from this conditional posterior distribution for $x_1^{mis(1)}$ given $\theta_1 = \theta_1^*$. This process is continued to then impute for $x_2^{mis(1)}$.

$$\theta_2^{(1)} \sim f(\theta_2).f(x_2^{obs}|x_1^{obs(1)}.\, x_3^{obs}.\, y.\, \theta_2)$$

$$x_2^{mis(1)} \sim f(x_2^{mis}|\, x_1^{obs(1)}.\, x_3^{obs}.\, y.\, \theta_2^{(1)})$$

The underlying imputation model for $x_2^{mis(1)}$ is as follows:

$$x_2^{mis(1)} = \theta_{20}^{(1)} + \theta_{21}^{(1)} x_1^{obs(1)} + \theta_{22}^{(1)} x_3^{obs} + \theta_{23}^{(1)} y + e_2$$

In the above statements, we observe the introduction of a new variable, $x_1^{obs(1)}$. This new variable now represents the imputed values of $x_1^{mis(1)}$ which contribute to the pool of observed data used to calculate the posterior distribution for θ_2.

$$\theta_3^{(1)} \sim f(\theta_3).f(x_3^{obs}|x_1^{obs(1)}.\, x_2^{obs(1)}.\, y.\, \theta_3)$$

$$x_3^{mis(1)} \sim f(x_3^{mis}|\, x_1^{obs(1)}.\, x_2^{obs(1)}.\, y.\, \theta_3^{(1)})$$

The underlying imputation model for $x_3^{mis(1)}$ is as follows:

$$x_3^{mis(1)} = \theta_{30}^{(1)} + \theta_{31}^{(1)} x_1^{obs(1)} + \theta_{32}^{(1)} x_2^{obs(1)} + \theta_{33}^{(1)} y + e_3$$

This process is continued till all missing values across all the variables in the dataset are imputed. This is the first iteration.

Generalising this to iteration (t) and variable (j),
$\theta_j^{(t)} \sim f(\theta_j).f(x_j^{obs(t-1)}|x - j^{obs(t)}\ y.\ \theta_j)$, where $(j = 1, 2, ..p)$.

$x_j^{mis(t)} \sim f(x_j^{mis}|\ x - j^{obs(t)-1}.\ y.\ \theta_j^{(t)})$, where $(j = 1, 2, ..p)$.
Advantages of MICE:- The approach is attractive as it does not require the conditional distributions to be normal in nature. This allows the method to be extended to using logistic regression for binary variable and ordered logistic regression for ordinal variables (Lee & Carlin, 2010). In addition, the method also allows the researcher to account for the complexities observed in the data (interactions, skip patterns, etc.) in the imputation model (van Buuren, 2007).

Disadvantages of MICE:- Although this method seems appealing, MICE is not without its limitations. The method is based on the assumption that the data is missing at random (MAR). Secondly, each conditional distribution needs to be specified separately. This would result in substantial modelling especially for datasets with many partially observed covariates. Due to this, the technique is more computationally challenging compared to joint modelling. This is because MICE splits the specification of a joint model into a series of univariate model specifications. The technique is more computationally challenging compared to joint modelling explained below. A key disadvantage of the multiple imputation via chained equations (MICE) procedure is that it is not grounded in theory like the joint modelling (JoMo) approach, and the conditional distributions may not always correspond to a joint distribution.

2.3.4 Joint Modelling

Joint modelling (JoMo) begins at the assumption that the data can be described by any multivariate distribution. Imputations are created as draws from the fitted distribution when ignorability (see section 2.1.4) is assumed. The standard model assumes all independent variables to be complete. Alternatively, we can restructure the model by placing all partially observed variables in the model as target variables in the imputation model. (Schafer & Yucel, 2002).

How does JoMo work?

The main idea of a joint imputation model if to define a multivariate joint model for all variables in the dataset for imputation of missing values in the outcome. Outcome here refers to the variables in the model with missing values and not the outcome of the analysis model for the substantive question. We recall that in this thesis, we refer to variables with the subscripts i,j and k to signal to the level at which the variable is captured. For example, X_{ijk} is a variable captured at the lowest level (Level 1) where (i) refers to the number of units at the lowest level, (j) refers to the second level (Level 2) and (k) refers to the third level (Level 3).
Now, suppose we have variables $X_{1,ij}$ and $X_{2,ij}$ are partially observed and $X_{3,ij}$ and$X_{4,ij}$ are variables with no missing data, then the simplest joint model is the multivariate normal model given by:

$$X_{1,ij} = \beta_{01} + \beta_{11}X_{3,ij} + \beta_{21}X_{4,ij} + e_{1,ij} + u_{1,j}$$

$$X_{2,ij} = \beta_{02} + \beta_{12}X_{3,ij} + \beta_{22}X_{4,ij} + e_{2,ij} + u_{2,j}$$

$$\begin{pmatrix} e_{1,i} \\ e_{2,i} \end{pmatrix} \sim N(0, \Omega_e)$$

$$\begin{pmatrix} u_{1,j} \\ u_{2,j} \end{pmatrix} \sim N(0, \Omega_u)$$

After defining the imputation model, Bayesian methods are used to fit the model and impute the missing data. Several models and algorithms can be used to impute missing values; Schafer (1997) proposed a joint multivariate normal model, fitted by MCMC. The main assumption of this method is that the partially observed data follow a joint multivariate normal distributionae5 (Quartagno & Carpenter, 2016). JoMo uses the Gibbs sampling approach by consistently drawing new values for all parameters i.e. the fixed effects (β), the covariance matrix (Ω) and the missing data. The current draw of missing values is combined with the observed data to make the first imputed dataset (Quartagno & Carpenter, 2019b). The sampler is run for sufficient iterations to guarantee independence between imputed datasets. Before running the Gibbs sampler, we also choose starting values for the parameters. While there is no reported difference in the results depending on the choice of these parameters, a more plausible choice of starting values would reduce the number of iterations and result in faster convergence of the sampler. The default choice is a matrix of $0s$ for fixed effects and an identity matrix for the Covariance matrix.

Advantages of JoMo:- One of the advantages of the JoMo is that it can be naturally extended to multilevel data structures. Such structures arise commonly, for example, when we have nested datasets or repeated observations. A Gibbs sampler approach is used within the method, that allows us to choose starting values for the parameters; the more plausible these starting values are, the faster the sampler converges. More general random effects structures can be modelled as well.

Quartagno and Carpenter (2019a) explored the validity and practical applications of the joint modelling approach for imputing missing values in continuous and categorical data. Neglible bias was observed in the fixed effects after imputation using the Joint Modelling approach, and the standard errors were smaller than what was observed in Complete Case Analysis (Quartagno & Carpenter, 2019a). Quartagno and Carpenter also concluded general latent normal variables algorithm worked very well for ordinal data, with negligible loss of efficiency.

Disadvantages of JoMo:- Considering a joint model on variables subject to missingness may not always be feasible or even realistic. For example, consider a survey with items targeted at different sub-populations; for example, item asking respondents when was their last pap smear or asking respondents the number of cigarettes smoked in the last 24 hours. This could apply to also questionnaires with a skip pattern. Imposing a joint distribution when a joint distribution may not even exist is not practicable. Secondly, a joint modelling strategy may not work when variables are nominal, count or semi-continuous variables (Yucel, 2008). Researchers must remain cautious when choosing the right method of imputation, bearing these factors in mind. The next section lists key factors to bear in mind, when defining the imputation model.

2.3.5 Choosing the Best Imputation Model

There are choices that a researcher needs to make before specifying the imputation model (van Buuren et al., 2015). It is important that one be aware of these decisions before proceeding to impute. These choices include:

1. **Mechanisms of Missingness**
   The researcher has to first identify if the data missing at random (MAR) or not (MNAR). Multiple imputation can be used under the assumption of MAR. The R package *mice* has been developed to accommodate data that is MNAR also though it requires some additional complex modelling. MNAR modelling always requires additional information from the user about how the distribution of the missing values differs from that predicted assuming MAR. There is limited knowledge on implementing MNAR within *jomo*.

2. **Structure of the Imputation Model** This includes both the structure of the dataset i.e. the imputation model should account for any dependencies or clustering in the dataset, as well as the residual distribution. The structure of the dataset should be reflected in the imputation and the analysis model. For example,imputation of missing values due to incomplete repeated measures, or missing values in higher level units, would demand the use of multilevel

regression model. In addition, the imputation model should also incorporate the relationship between the target variable i.e. the variable with missing values and its predictors.

3. **Selecting Predictors for the Imputation Model**
   While the general advice is to include as many variables and their interactions in the imputation model as possible as recommended by both van Buuren(2018) and von Hippel (2009), we need to be sure whether the predictor variables actually predict the target variable or not. In it imported that we ensure that the analysis model is a subset of the imputation model. A misspecified imputation model is worse than failing to impute for the missing values. This is discussed in greater detail in Section 2.3.7 below.

4. **Imputing Derived Variables**
   This refers to handling variables that are a function of the variable being imputed - also known as the target variables. Researchers would often encounter variables that are derived from the target variables. These include linear and non-linear transformations. Imputing derived variables can be done by either the **Transform then Impute** method, or the **Impute then Transform** method (von Hippel, 2009). Regardless of the method used, the researcher must ensure that the relationships between variables is maintained post imputation. This is discussed further in section 2.3.9 below.

5. **Order of Imputation**
   This decision depends on the pattern of missing data observed in the dataset. If a monotone missing data pattern is observed then ordering the variables in an increasing order of missing values would help speed convergence (van Buuren et al., 2015). A variable can be visited multiple times in the same iteration - this is applicable to both MICE and jomo. This would help improve synchronisation between the variables, as it ensures that any partially observed covariate has been imputed prior to its occurrence in the imputation model, and therefore is part of the observed data i.e. X^{obs} . This becomes more relevant when imputation is sequential such as in MICE.

A key attraction of MI is that any variable that is not included in the analysis model can still be used in the imputation model. We recall the definition of MAR i.e. the probability of missingness is conditionally independent of the observed data. Thus including carefully selected auxiliary variables in the imputation model may improve the plausibility of the MAR assumption. This is demonstrated in the following chapters by the inclusion of variables X_{2ijk} and Z_{2jk} in the imputation models for X_{ijk} and Z_{jk}; but not in the analysis model (see Table 4.2.2).

2.3.6 Is there an ideal number of imputations?

The basic idea behind multiple imputations is to replace every missing observations with several say m plausible values. This would account for the uncertainty that is ignored during single imputation. These m imputations create m complete datasets that are analysed using standard complete-data techniques. At such a time we can ask the question, what is the ideal value of m and does m change with the degree of missingness observed in the data?

Rubin (1987) states that using a value of m that lies between say 2 and 10 is relatively efficient for a modest fraction of missing information. He also showed the efficiency of finite-m repeated imputation estimator relative to the infinite-m repeated imputation estimator is $\frac{1+\gamma_0}{m} - 1/2$, where γ_0 is the population fraction of missing information. Rubin further illustrated that across different fractions of missing information (FMI), for m lying between 2 and 5, the large sample coverage probability remains constant. This explains why the default number of imputations in most statistical packages is set to 5. This has also resulted in the guideline that advises that more than 10 imputations are rarely required.

Bodner (2008) questioned the impact of this guideline on the p-values, confidence interval half widths and estimated FMI. He demonstrated that variability of estimates

decreases substantially as number of imputations increases. Extending Bodner's concerns to a multilevelled scenario, we can question if the number of imputations have an impact on the estimates across

- Different proportions of missingness
- Different sample sizes
- Different cluster sizes
- Different mechanisms of missingness

White, Royston, and Wood (2011) argue that efficiency and reproducibility of the analysis increases with the an increased number of imputations, as Monte Carlo error tends to zero as number of imputations increases. In agreement with Bonder's rule, White et al. suggested requiring about $m \geq 100 * FMI$. In instances where FMI is unknown, they recommend using *"Bodner's approximation that the FMI for any parameter is likely to be less than the fraction of incomplete cases"*.

In this thesis, we have accepted Rubin's general rule of having the number of imputations maintained between 5-10. We observed increasing the number of imputations had minimal impact on the standard errors of the estimates obtained and increased computation time in the simulation study in Chapter 5. The next section discusses the pressing topic of how to develop the imputation model.

2.3.7 Identifying Predictors for the Imputation Model

Murray et al. (2018) provides three recommendations when developing an imputation model :

- *Imputations should reflect uncertainty about the imputation model.*

  Suppose we have a single covariate with missing values to be imputed using a regression model. If the objective was to only "fill in" these missing values, we would just impute using the fitted values. However, the theory of multiple

imputation is based on incorporating for the uncertainty that missing values brings to the dataset via some stochastic mechanism. We also need to account for the uncertainty from the imputation model itself. Methods that do not appropriately reflect both sources of uncertainty can underestimate the between-imputation variance, yielding small standard errors and anti-conservative inferences (Rubin, 1996).

- *Imputation models should include as many variables as possible.*

  The MAR assumption tends to become more plausible as more complete variables are added into the imputation model. In addition, as Rubin stated in (1996), *"if predictors of the covariates with missing information are not added into the imputation model, but are added to compute the estimand of interest and the sampling variance, then the imputations will be improper and further result in biased estimate due to repeated imputations"*. To summarise, the price of excluding relevant predictors from the imputation model is much higher than the cost of including irrelevant variables in the model. This is important in situations where the imputer and the analyst are not the same person, and the imputations must be compatible with many unspecified analyses. Even in circumstances where the imputer and the analyst are the same person, there should be certain standardization of the imputation procedure that supports all analysis rather than generating a new set of imputations for every analysis performed. Reiter, Raghunathan, and Kinney (2006) further emphasised the importance of accounting for the sampling design in imputations to achieve valid inferences. Sampling designs can include specifying the cluster identifying variables in the imputation process, and including clusters as fixed or random effects in the imputation model.

- *Imputation models should be as flexible as possible.*

Imputation models should "track the data" (Rubin, 1996) by modelling the relevant aspects of the joint distributions of the missing data. This could include interactions, non-linear transformation of variables, etc. This ties back to the recommendation that the imputation model should adopt the sampling design of the study and the structure of the data. Despite their stochastic nature, the imputations should be robust across changing imputers and analysts. This further highlights the need to develop congenial and robust imputation models that are not subjective to the imputer.

Influx and Outflux

In his book, van Buuren (2018) proposed using Influx and Outflux statistics to streamline this process of constructing imputation models that are impartial to the subjectivity of the imputer and the analyst. These statistics quantity how each variable connects to others. As described by van Buuren (2018), for a pair of variables (X_a, X_b) in a sample of n observations with p variables, the influx coefficient for the variable X_a is defined as

$$I_a = \frac{\sum_{a=1}^{p}\sum_{b=1}^{p}\sum_{i=1}^{n}(1 - r_{i,a})r_{i,b}}{\sum_{b=1}^{p}\sum_{i=1}^{n} r_{i,b}}$$

The value of I_a depends on the proportion of missing data in the variable X_a. This means that if X_a has no missing values, $I_a = 1$. The influx coefficient (I_a) is the proportion of variable pairs (X_a, X_b), where X_a is missing and X_b is observed. Thus, a high influx implies how well the variable is connected with the observed data.

The second measure proposed by van Buuren (2018), is the outflux coefficient. For the same pair of variables (X_a, X_b), the outflux coefficient indicated how useful is X_a to impute missing values in X_b. Unlike the former, outflux depends on the extent of missing values in the variable X_a. The variable with higher outflux is better connected to the missing data, and can be more useful in imputing other variables. The outflux coefficient is given by

$$O_a = \frac{\sum_{a=1}^{p}\sum_{b=1}^{p}\sum_{i=1}^{n} r_{i,a}(1 - r_{i,b})}{\sum_{b=1}^{p}\sum_{i=1}^{n}(1 - r_{i,a})}$$

Here $r_{i;a}$ and $r_{i;b}$ are response values (of 1s and 0s) observed for the i^{th} observation for the variable X_a and X_b respectively.

Fluxplots can be used to plot both these measures to make interpretation easier for the imputer and the analyst. It can be used to identify variables that clutter the imputation model, thus making the process of constructing the imputation model less subjective. Variables in the lower areas in the fluxplot that are not used for analysis can be removed from the imputation model. While there is a significant proportion of literature that focuses on developing proper imputation models, there is an evident lack of illustrative examples describing the imputation process in practice to create imputation models which are impartial to the imputer and the analysts and relies on quantitative measures to develop these models. Case Study on Hepatitis C in Section 7.4 succeeds in illustrating the imputation model development process using Influx and Outflux in the context of missing values in laboratory data.

2.3.8 Uncongenial Imputation and Analysis Models

Before we begin discussing the relevance of having imputation and analysis models that are congenial, we need to understand what uncongenial means. In simple words, uncongenial means the imputation and analysis models make different assumptions about the data. It typically arises when both the analyst and the imputer have access to different sources of information. As Meng (1994) stated *"If the imputer's model is reasonably accurate, then following the multiple-imputation recipe prevents the analyst from producing inferences with serious non-response biases"*.

Xie and Meng (2017) definition of congeniality as eloquently explained by Jonathan Bartlett[1]is that there exists a Bayesian model for the data such that:

- given complete data, the point estimate given by fitting our analysis model

[1]Bartlett J 2020, Auxiliary variables and congeniality in multiple imputation, The Stats Geek, accessed 5 January 2020, <https://thestatsgeek.com/2020/12/05/auxiliary-variables-and-congeniality-in-multiple-imputation/>

of interest to that data matches the posterior mean of the parameter of interest, and the variance estimator calculated by our analysis model matches the posterior variance.

- the conditional distribution of the missing data given the observed in this Bayesian model matches that used by our imputation model.

If they are congenial and the models are correctly specified, Rubin's variance estimator is (asymptotically) unbiased for the true repeated sampling variance of the MI point estimator(s). Uncongeniality may occur in situations where the imputation and analysis models make different assumptions about the data generating processes and are thus uncongenial. Bondarenko and Raghunathan (2016) show that there are different impacts on the validity of the imputations depending on the type of uncongeniality observed. For instance, if the analysis model is a subset of the imputation model, it will typically result in unbiased point estimates with wider confidence intervals. In the reversed scenario, where the imputation model is a subset of the analysis model, one would typically observe biased point estimates.

Another occurrence of uncongenialty is when the model assumptions are consistent for both the imputation and analysis models, but the analyst prefers a suboptimal analysis procedure (e.g. method of moments over MLE). The consequences of using clusters as fixed effects in the imputation model when the analysis model is based on random effects is another example of uncongeniality(Reiter et al., 2006).

This raises the question of how to avoid developing uncongenial models. Techniques like using standard regression diagnostics, posterior predictive checks and variable selection methods all help (Bondarenko & Raghunathan, 2016). These diagnostic methods help understand data structures and model features, allowing the imputer to integrate them into the imputation model.

2.3.9 Imputing Derived Variables

Reiter et al. (2006) released a paper on the importance of incorporating the sampling design into the imputation model. This ties back to the importance of congeniality between the analysis and imputation model, not only in terms of the predictor variables, but also the sampling design employed. This means including strata and/or clusters as fixed or random effects, so as to reflect the structure of the data in the imputation model. Reiter et al. illustrated the severity of bias that occurs when imputers fail to account for the complex study designs in the imputation process, and also suggested approaches to account for design features into the imputation model.

Among their suggested approaches was to use hierarchical models where effects of clusters are incorporated as random effects. While their suggestions have been taken on board by researchers dealing with clustered data, there is still several gaps that need to be addressed. Among these is the primary issue of derived variables.

Derived variables are a commonly occurring phenomenon in statistical analysis. These include linear and non linear transformations such as mean, squares, square roots, logarithmic transformations, interactions between variables etc. Derived variables are also a natural consequence of analysing clustered data. Often our analysis requires calculation of cluster means and deviation of variables from their cluster means. Regression analysis with missing values gets trickier when the regressions include these derived variables. ***"Passive Imputation"*** was recommended by von Hippel (2009) to handle these derived variables. In this method, the imputer develops an imputation model for the variable with missing values, imputes this variable in its original form and then transforms it using the imputed values. This can be referred to as the ***Impute, then Transform*** approach.

An alternate method to handle derived variables is to transform the variable with missing data, and then impute the transformed variable like any other partially observed covariate in the dataset. This method follows a ***Transform, then Impute*** ap-

proach and is referred to as the ***"Just Another Variable (JAV)"*** approach. Both Passive Imputation and JAV approach were recommended by von Hippel (2009) as they yield good regression estimates. The disadvantage of using the JAV approach is the disconnect between the original values and the imputed values. Since the transformed variable is imputed as any other partially observed variable, the imputed values of the transformed variables will not be a direct transformation of the original variable. The study (von Hippel, 2009) was limited to two types of transformations i.e. squares and interactions. The study calls for future research looking at the validity of both methods in more complex scenarios including multilevel data.

Seaman, Bartlett, and White (2012) investigated imputation of incomplete covariates when the model of interest includes more than one function of that variable as a covariate, focussing on the methods proposed by von Hippel (2009). In their study, Seaman et al. identified that when data are MAR, all the three methods i.e. JAV, Passive Imputation and PMM provided biased estimates. JAV performed poorly when a logistic regression model was used for analysis.Seaman et al. recommended using JAV for linear regression with a quadratic or interaction effect.

2.4 Summary

In this chapter, we introduced key concepts and common terminology that are used within the field of missing data. We also introduced the different mechanisms of missingness as well as the concept of multiple imputation. Within multiple imputation, we introduced popular methods of MI i.e. MICE and JoMo and highlighted their functions, strengths and limitations. Finally we threw light on several key aspects one must consider before performing multiple imputation, such as the number of imputations, order of imputations, congeniality and passive imputation. This chapter is crucial as it sets the tone for the thesis. Understanding how these methods work is critical as they are a recurring theme within the thesis.

The thesis titled *"Multiple Imputation of Missing Data in Multilevel Models"*, has two key components - 1) Multiple Imputation and 2) Multilevel Models. Thus this thesis highlights two aspects of data i.e. quality of data collected and the structure of the data which often goes back to the sampling design used for data collection. In this chapter, we introduced the first part of the jigsaw i.e. the various characteristics of missing data, and introduced methods of multiple imputation i.e. MICE and JoMo.

In the next chapter, we look at the second part of the jigsaw - multilevel models in particular and the intersection between missing data and multilevel models that forms the focus of this thesis- ***multilevel multiple imputation.***

Chapter 3

Multilevel Multiple Imputation and Related Work

In this chapter, we define what we mean by multilevel models, fixed and random effects, we give a mathematical definition of different types of multilevel models and examine how information is captured at each level. Finally we describe the literature around multiple imputation in multilevel models. This chapter is important as it provides the necessary information about the structure of the data that will be used throughout the thesis and provides the second part of the jigsaw i.e. **multilevel models**. Section 3.1 provides an in-depth understanding of the concept and structure of multilevel models. In this section, we formulate the structure of a three-level model for the first time (see subsection 3.3). This is the structure of the model that will be used in defining the simulation study. In Section 3.4, we discuss the existing literature on the intersection between multiple imputation and multilevel models and highlight the ideas that we have inherited and extended in this thesis. This section is crucial as it provides the landscape within which we can place this research and also helps to identify gaps in the current body of literature. Figure 3.4.2 helps illustrate the current landscape of the body of knowledge surrounding multilevel multiple imputation. It also helps the reader identify the gaps in the current literature, and the extensions of existing concepts that the thesis has taken.

3.1 Multilevel Models

Multilevel modelling is a statistical technique that extends ordinary regression analysis to the situation where the data is hierarchical. Alternate terms used for multilevel modelling include hierarchical linear models, hierarchical modelling, random effects models and variance heterogeneity models. It seeks to identify the effect of both cluster and individual level variables on the individual level outcome. This type of analysis takes into account the positive (or negative) correlations within clusters which if ignored would overestimate the standard deviation of the effect of cluster variance. Most health data sets are multilevel in nature. For example, clustering can be induced by patients sharing the same environment and nursing resources in a hospital. Invoking the independence assumption for such data would lead to invalid inferences as the precision for the effects are often over or under estimated. This assumption might also result in atomistic fallacy that is, drawing inferences regarding the variability across groups based on individual data (Diya, 2011).

Higher level units in a multileveled structure are important because they define the action space of individuals. Many problems in public health policy are related to people's behaviour. People behave within the social and institutional context of their community or workplace. The general idea of multilevel analysis is that this hierarchy is taken into account in the analysis, or in other words, the analysis takes into account the dependency of the observations (Twisk, 2006).

3.1.1 Fixed versus Random Effects

Fixed and random effects are concepts commonly used, especially in experimental research where different treatment groups are involved. A factor defining various treatments is said to have a fixed effect if all possible treatments in which the researcher is interested are present in the experiment. A random effect is attributed to a factor defining treatments that can be considered a sample from the population of all relevant treatments (Gelman & Hill, 2006).

Consider another example; wherein we wish to test the effectiveness of an intervention programme. The researcher uses already existing groups such as schools to administer their intervention. These groups are not alike and are a random sample from all the possible schools in the population. The effect of the school on the intervention program varies across schools and forms a random factor.

The distinction between fixed and random effects is important as it has consequences on the inference that can be made, and on the generalizations of the results. Fixed effects can only allow inferences to be made regarding the intervention that is used in the experiment. These effects are assumed to be constant and without any measurement error. For a random effect, the inferences are extended beyond the the levels observed (e.g. schools) in the sample. The objective is to generalize the results of the intervention to all the schools in the population and not just the ones in the sample selected. Thus, these effects are not assumed to be constant, but from a distribution or measured with sampling error (Kreft & De Leeuw, 1998; Gelman & Hill, 2006). Model specification of multilevel data with random effects is elaborated on in the next sections.

3.2 Model Specification

Again, we recall that we refer to variables with the subscripts i,j and k to signal to the level at which the variable is captured. X_{ijk} is a variable captured at the lowest level (Level 1) where (i) refers to the number of units at the lowest level, (j) refers to the second level (Level 2) and (k) refers to the third level (Level 3).
To provide clarity to the reader, the model is introduced in the context of a specific survey problem involving individuals (Level 1) nested within households (Level 2) that are further nested within districts (Level 3). Here, the subscript (i) denotes the individuals, (j) denotes the households and (k) denotes the districts. Readers should note that throughout this thesis, the covariates measured at the lowest level are signalled to by the letter X, the covariates at the second level are denoted by the letter Z and the covariates at the highest levels are denoted by the letter W. Outcome vari-

ables are denoted by the letter Y.

Assume Y_{ij} is measured for the i^{th} individual in the j^{th} cluster, where there are (n_j) units in the first level and (J) units in the second level. The simplest of the hierarchical linear models is the equivalent of the ANOVA model with random effects.

$$Y_{ij} = \beta_{0j} + e_{ij}; \tag{2.1}$$

where the error term follows $N(0, \sigma^2)$. Here, $E(\beta_{0j}) = \gamma_{00}$, and $Var(\beta_{0j}) = \tau_{00}$. This model is referred to as a fully unconditional model (Raudenbush & Bryk, 2002) as it contains no predictors either at level 1 or level 2. Here Y_{ij} is predicted using a level 2 parameter namely β_{0j} which varies across each level 2 unit and can be interpreted as the mean outcome for the j^{th} unit.
β_{0j} is further modelled as

$$\beta_{0j} = \gamma_{00} + r_{0j} \tag{2.2}$$

where γ_{00} represents the grand mean outcome in the population and r_{0j} is the random effect associated with the j^{th} unit with $r_{0j} \sim N(0, \tau_{00})$.
Combining equations (2.1) and (2.2), we get

$$Y_{ij} = \gamma_{00} + r_{0j} + e_{ij} \tag{2.3}$$

which in comparison is the same as the ANOVA model with group effect (r_{0j}) and individual effect(e_{ij}). This is a random effects model as the group effects are random. It can be shown that,

$$Var(Y_{ij}) = Var(\beta_{0j} + e_{ij}) = \tau_{00} + \sigma^2.$$

It can be highlighted that the within group variability is captured by σ^2 and the between group variability is captured by τ_{00}. This relationship can be used to estimate

the intra cluster correlation coefficient (ICC) denoted by (ρ) (Menon, Bhaskarapillai, & Richardson, 2019). The ICC is a measure of relatedness within clustered data and is calculated as $\rho = \tau_{00}/(\tau_{00} + \sigma^2)$.

Gelman and Hill (2006) stated that multilevel models can be thought of sitting between two extremes that are available to researchers when clustering is observed. These are referred to as *fully pooled* and *fully unpooled* models (Gill & Womack, 2013). The fully pooled model treats the group level variable as an individual variable implying that the group level distinctions are ignored. This will now be explained.
For example, if X_{ij} is a variable measured at individual level and Z_j is a variable measured at group level where there are n_j units at the first level and J units at the second level, then

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + e_{ij}$$

We note, that unlike equation(2.3), there is no r_{0j} specified, indicating that the group effects do not matter and that the cases should be treated homogeneously ignoring any variation between groups.

At the opposite end of the spectrum, we have the *fully unpooled* model in which we treat each group as a separate dataset and then model them separately. This is shown below,

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$$

Unlike the former, the fully unpooled model assumes heterogeneity as each case belongs to one of the (j) groups, $j = 1, 2 \ldots J$. Here, we improve the model fit by classifying each i^{th} unit into its respective group by loosening the definition of the single intercept i.e. β_{0j} which gives the n_j observations a common intercept if they fall in the same group. We see that the variable Z_j does not enter the model as it is modelled as a sub part of the intercept. Here there are no levels indicating hierarchy and β_{0j} and β_{1j} are assumed to be fixed parameters unlike the distributional assumptions made in random effects models.

The fully unpooled approach is starkly different from the fully pooled approach as it highlights that the groups are so fundamentally different that they cannot be modelled together. Between these two approaches lies the multilevel modelling approach where the clusters are recognised as different but are associated by common individual level fixed effects and also have distributional assumptions made on the random effects (Gill & Womack, 2013).

Prior to the development of multilevel analysis, there were two options for analysing the data from different levels with ordinary least square regression. The data would be aggregated to a higher level, which would result in the *ecological fallacy*. This would result in loss of information and misinterpretation of results. Another technique used would be to distribute the characteristics of higher levels to all individuals which would result in the *atomistic fallacy* also known as the *individualistic fallacy*.

Adding in level 1 and level 2 covariates to the model specified in equation (2.1), we get

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \tag{2.4}$$

The level 2 model becomes,

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + r_{0j} \tag{2.5}$$

$$\beta_{1j} = \gamma_{10}. \tag{2.6}$$

Substituting equations (2.5) and (2.6) in (2.4) , we get

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + r_{0j} + e_{ij} \tag{2.7}$$

Here γ_{00} is the average intercept across level 2 units, γ_{10} is the average regression slope across level 2 units,and r_{0j} is the increment to the intercept associated with each level 2 unit (j). Here, by setting $\beta_{1j} = \gamma_{10}$, we also assume that the covariate

effect is identical across level 2 units.

3.2.1 Centering of Covariates around the Group Mean

In regression modelling, it is important for the variables in the model to have a precise meaning so that statistical results can be related to theoretical interests in the research. For example, in hierarchical modelling, β_{0j} can be interpreted as the expected outcome for the i^{th} individual in the j^{th} level 2 unit, when the value of $X_{ij} = 0$. In most cases, X_{ij} taking value zero is not meaningful (example, age, height, weight, etc.). In such situations, the researcher may want to choose a location for X_{ij} such that β_{0j} can be interpreted in a more meaningful way.

Raudenbush and Bryk (2002) defined 3 such possible locations of X; *'natural X metric'*, *'centering around the grand mean'* and *'centering around the group mean'*. Centered variables are those whose values have been reset as deviations from a particular value. From equation (2.7) , we note that the variable X_{ij} is centered around its group mean, i.e. $X_{ij} - \bar{X}_j$. This is also referred to as *'centering within context'*. Centering variables in multilevel modelling has proven to be helpful in terms of providing stability and computational ease (Kreft, De Leeuw, & Aiken, 1995). Centering in multilevel models can also be useful in separating individual level effects from group level effects.

3.2.2 Post Imputation Centering

"In Complete Case Analysis, centering provides a pick-a-point approach when estimating slopes" (Enders, Baraldi, & Cham, 2014). Enders et al. explain that when meaningful cut-offs are unavailable, researchers often prefer to interpret conditional effects at different points in the normal distribution. In such cases, centering of covariates before imputation would require a new imputation model to be defined for each value of $X_{ij} - \bar{X}_j$ calculated. To avoid this issue, they suggested post imputation centering of variables. As in the case of derived variables (see section 2.3.9), the method involves imputing the variables on their original scale,rescaling them post imputation, and obtaining the relevant estimates from the regression model.

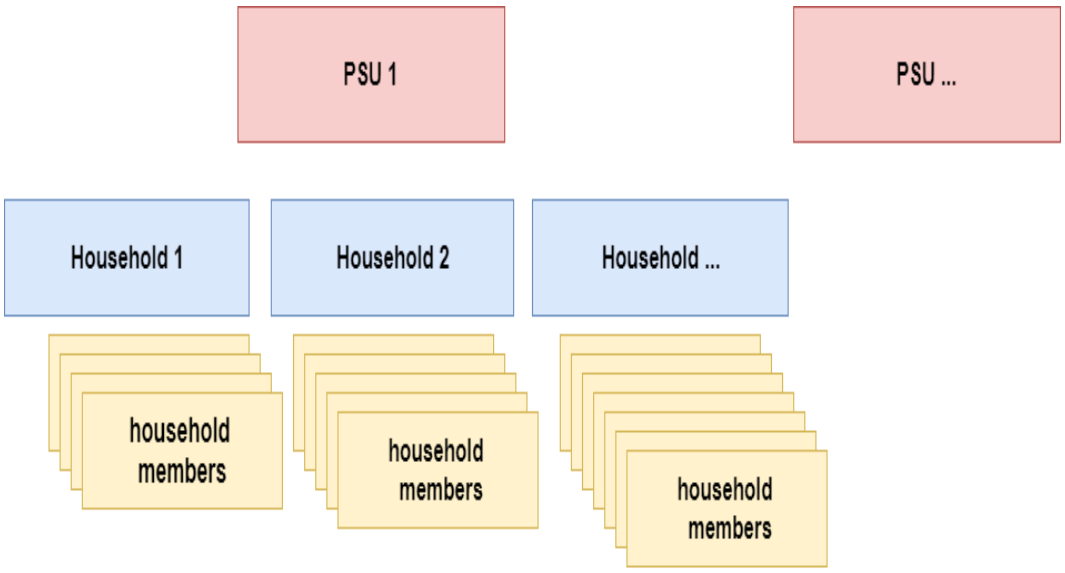


Figure 3.3.1: Illustration of a three-level hierarchical data structure

Centering for higher level predictors are given as $X_i^{dev} = X_i - \bar{X}$; and $X_{ij}^{dev} = X_{ij} - \bar{X}_j$. We employed post imputation centering methods in our simulation study. While the choice of location of level 2 predictors are not crucial in simulation, researchers have shown that it does provide more stability and easy interpretation of coefficients. It is thus recommended that Level 2 predictors are centred around their corresponding grand mean (Enders et al., 2014).

3.3 Formulating Three Level Models

The same logic can be extended to formulate a three-level model. To provide clarity to the reader, the model is introduced in the context of a specific survey problem involving individuals (Level 1) nested within households (Level 2) that are further nested within districts (Level 3). Here, the subscript (i) denotes the individuals, (j) denotes the households and (k) denotes the districts. Again, we remind the reader that, we refer to variables with the subscripts i,j and k to signal to the level at which the variable is captured. Here, the subscript (i) denotes the individuals, (j) denotes the households and (k) denotes the districts. Readers should note that throughout

this thesis, the covariates measured at the lowest level are signalled to by the letter X, the covariates at the second level are denoted by the letter Z and the covariates at the highest levels are denoted by the letter W. Outcome variables are denoted by the letter Y.

Bearing that in mind, the variable Y_{ijk} can be interpreted as the variable measured for the i^{th} individual in the j^{th} household residing in the k^{th} district (Menon et al., 2019).

The simplest three-level model i.e. the fully unconditional model can be specified as

$$Y_{ijk} = \beta_{0jk} + e_{ijk} \tag{2.7}$$

$$\beta_{0jk} = \gamma_{00k} + r_{0jk} \tag{2.8}$$

$$\gamma_{00k} = \delta_{000} + u_{00k}. \tag{2.9}$$

Thus,

$$Y_{ijk} = \delta_{000} + e_{ijk} + r_{0jk} + u_{00k} \tag{2.10}$$

where, $i = 1, 2 \ldots n_{jk}, j = 1, 2 \ldots J$ and $k = 1, 2, ...K$.

Now, model (2.8) is the household level model where we view each household mean (β_{0jk}) to be varying randomly around some district mean (γ_{00k}). Similarly, the model (2.9) is the district level model where each district mean (γ_{00k}) varies randomly around a grand mean (δ_{000}) (Menon et al., 2019).

The model (2.10) represents how the total variation in the outcome variable is allocated across three levels i.e. Level 1 is individuals within households, (σ^2); Level 2 is households within districts (τ_r); and Level 3 is for among districts (τ_u); thus allowing the estimation of the proportion of variance at each level (Menon et al., 2019).
This implies that $\frac{\sigma^2}{(\sigma^2 + \tau_r + \tau_u)}$ is the proportion of variance among individual; and

$\frac{\tau_r}{(\sigma^2 + \tau_r + \tau_u)}$ is the proportion of variance among individuals within households; and

$\frac{\tau_u}{(\sigma^2 + \tau_r + \tau_u)}$ is the proportion of variance among districts.

Equations (2.7) - (2.10) are the unconditional model and allow for estimating variability that is associated with the levels in the model. However, variability in an outcome variable can be due to the variables that are measured at each level in the model. Furthermore, we can also expect relationships between variables at different levels to vary randomly across units.
For example, suppose that within households, child's nutrition and growth is related their haemoglobin (Hb) concentrations. A study conducted in Sub-Saharan Africa (Moschovis et al., 2018) showed that the child's nutrition and growth is influenced by other risk factors such as household family size, wealth index & distance of water sources from the house.

Due to the complex interconnectedness of many of these risk factors measured across levels, it is important to model these risk factors appropriately. These possibilities encourage a general structure at each level (Raudenbush & Bryk, 2002, p. 178). Using the notations from Menon et al. (2019)

General Level 1 Model:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{1ijk} + \beta_{2jk}X_{2ijk} + ... + \beta_{pjk}X_{pjk} + e_{ijk}(i = 1, 2, ...p), (j = 1, 2, ...q), (k = 1, 2, ...s) \tag{2.11}$$

General Level 2 Model:

$$\beta_{pjk} = \gamma_{p0k} + \sum_{j=1}^{q}\sum_{k=1}^{s}\gamma_{pjk}Z_{pjk} + r_{pjk} \tag{2.12}$$

Here,γ_{p0k} is the corresponding district level coefficient that represents direction and strength of association between Z_{pjk} and β_{pjk}. $\beta + X_{pjk}$; and Z_{qjk} is the vector of

household level predictors for effect β_{0jk}. It can be noted that each β_p may have a unique set of household predictors Z_{qjk}. The corresponding coefficient representing direction and strength of association between Z_{qjk} and β_{pjk} is given by γ_{pqk}. Finally, $r_{pjk} \sim N(0, \tau_r)$.

$$Var(\tau_r) = \begin{bmatrix} \tau_{11} & & \\ & \ddots & \\ & & \tau_{pp} \end{bmatrix} \;\&\; Cov(r_{pjk}, r_{p'jk}) = \tau_{pp'}$$

General Level 3 Model:

$$\gamma_{pqk} = \delta_{pq0} + \sum_{k=1}^{s} \delta_{pqk} + W_k + u_{pqk} \tag{2.13}$$

where δ_{pqk} is the corresponding level 2 coefficient that represents direction and strength of association between W_k and γ_{pqk}. W_k is the level 3 characteristic used as a predictor for γ_{pqk}. As in equation (2.12), each γ_{pq} may have a unique set of W_k; $k = 1, 2....s$, and $u_{pqk} \sim N(0, \tau_u)$. In each of these models, the coefficients(including the intercept) can be either random of fixed. If the coefficients are fixed, then the corresponding errors are assumed to be zero.

3.4 Related Work

In this section, we will be reviewing the existing body of work on multilevel multiple imputation (MI). This section is crucial as we will be focusing on significant papers and work of the major contributors within the field of multilevel MI. These papers have provided the foundation on which my research rests. I draw on the knowledge that these studies provide as a way of strategically positioning my research to show where my contribution to this knowledge lies in this body of literature. Before undertaking any research, it is important to know the existing body of work and to be aware of what aspects of these works are inherited and departed from in this thesis. Figure 3.4.2 is an illustrative guide to help readers conceptualize the various dimensions in multilevel multiple imputations, and the gaps identified in the

existing literature. Multilevel Multiple Imputation is one of the current hotspots in statistical technology providing imputers and analysts an array of options while handling missing values in multilevel data. In the previous sections, we have addressed what the structure of a multilevel model looks like, and by extension the importance of accounting for the structure of the data in analysis during research. Before we address missing data in multilevel models, it is important to understand why multilevel structure and by extension, multilevel analysis is important in research. The general idea of multilevel analysis is that hierarchy is taken into account in the analysis or in other words, multilevel analysis takes into account the dependency of the observations (Twisk, 2006).

In single level models, missing values can occur in either the predictor variable or the outcome variable or both. Similarly, in multilevel models, missing values can occur in the outcome variables or the covariates or both. These covariates can be captured at any level in the data. It is also possible to have missing values in the cluster identifying variable (or class variable). However, this is beyond the scope of this thesis.

Previous studies have seen the use of single imputation for missing values in higher level variables by ignoring the multilevel structure. This resulted in underestimation of the standard error (Reiter et al., 2006). This has led to a body of work centred around imputing missing data when it occurs in higher level covariates (Enders, Mistler, & Keller, 2016; Huque et al., 2019; Audigier & Resche-Rigon, 2017).

Gelman and Hill (2006) formulated a statement on the imputation of missing data in multilevel models. This statement sets the foundation for this thesis. In a scenario with missing data in multilevel models, their advice was to *"create two different datasets for individual and group level data. The group level dataset would include aggregate forms of individual level measurements when imputing for missing values in this level"* (Gelman & Hill, 2006, p, 541). This statement primarily refers to a 2 level model. This approach was also explored by Lüdtke, Robitzsch, and Grund (2017) and Yucel (2008).

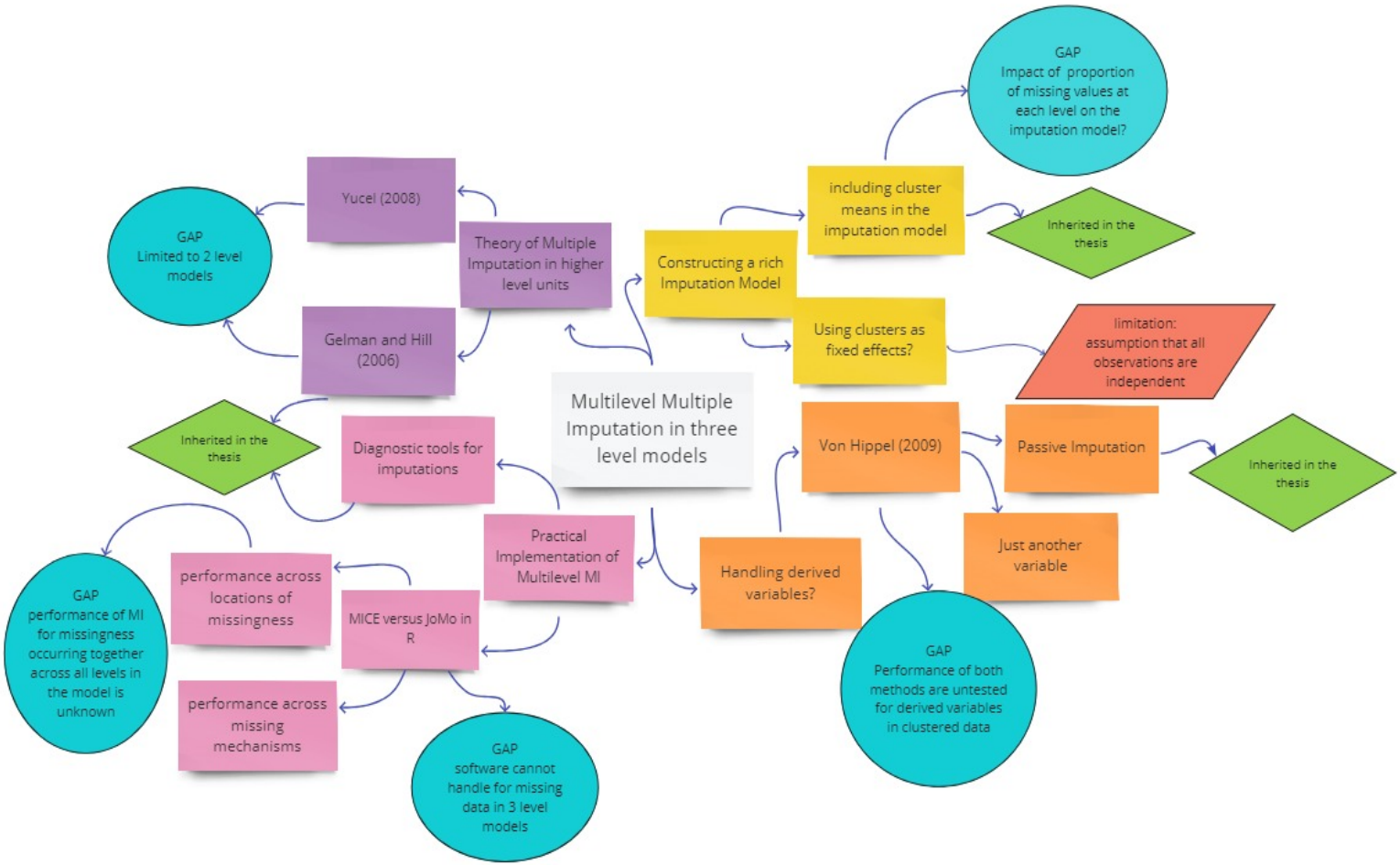


Figure 3.4.2: Conceptualization of dimensions and gaps in the literature on multilevel multiple imputation

Yucel (2008) stated that *"Despite many advances in both theory and applied methods for missing data, missing-data methods in multilevel applications lack equal development"*. This comment holds true to some extent even today. Although many would argue that MI methods have seen a recent boom in software implementation, multilevel multiple imputation has only received attention in the last five years.

Critical literature focusing on this aspect includes works by Enders et al.(2016), Huque et al. (2019), van Buuren (2007), Quartagno and Carpenter (2019b), and Lüdtke et al. (2017). Across this array of literature, we see that there still seems a gap in literature on both implementation and performance of multiple imputation for data with more than two levels, a gap in understanding performance of MI when there are additional covariates in the model, and how this performance changes with changing cluster sizes across levels in the data.

Yucel (2008) elaborated further on multivariate multilevel MI when missing values occur arbitrarily across any levels in the data. The study uses the Gelman & Hill theory of imputing missing values in each level separately, and jointly imputes missing values in variables measured at any level in the study. This method has been largely accepted across literature for imputing missing values in higher level units (Enders et al., 2016), and is also currently employed in the packages and tools across statistical software. The simplicity of the approach, and it's flexibility makes it highly attractive and a reasonable approach grounded and tested in theory for multilevel imputation (Lüdtke et al., 2017).

In addition, the current tools are limited to imputation for two level datasets only with the exception of BLIMP software developed by Enders et al. (2018). Existing imputation routines are diverse and offer different functionality; all approaches can accommodate basic random intercept analyses but they differ in their functionality to estimate random slopes,binary and ordinal variables, within cluster and between cluster covariance matrices, and partially observed covariates at higher levels (Enders

et al., 2018). BLIMP currently remains the only tool that successfully imputes missing values in the third level.

Yucel (2008) uses Level 1 characteristics in an *auxiliary* sense by aggregating these (imputed or observed) values to the next level. In his method, Yucel (2008) iterates between 3 interrelated Gibbs samplers, each approximating level 1,2 and 3 observed data posteriors and posterior predictive distributions (as described in Section 2.3.3). The algorithm is a step wise procedure starting with imputation at Level 1 that is run till the algorithm converges. Using imputed values from the previous step, missing values at Level 2 are imputed. The process can be extended in theory to Level 3 using a separate Gibbs sampler using quantities from the previous two steps. The paper drives the theory behind imputation of missing data in 3- level models in this thesis. Software since then has implemented these techniques their MI procedures at 2 or 3 levels.

The thesis implements the step wise iterative procedure similar to the procedure described by Yucel (2008). The thesis also incorporates aggregated values of lower level units that both Gelman and Hill, and Yucel have recognized as valuable to create a rich imputation model. A rich imputation model is one whose purpose is not to infer on the underlying population but to generate plausible imputations while preserving the structure of the dataset and accounting for the mechanisms of missingness. Yucel's work in 2008 on multilevel multiple imputation has provided the theoretical foundation upon which several studies that focus on multilevel imputation rests.

Several studies have discussed and compared different methods of multiple imputation and their performance in comparison to other methods of imputation (Kalaycioglu, Copas, King, & Omar, 2016; Schmitt, Mandel, & Guedj, 2015; Kropko, Goodrich, Gelman, & Hill, 2014). Imputation using Bayesian variable selection is widely discussed in the literature (Rubin, 1987; Little & Rubin, 2014; Yang, Belin, & Boscardin, 2005). Andridge (2011) addressed the impact of impact of using fixed effects for clusters in

the imputation model. Using simulations, Andridge compared the fixed effect imputation model to a model that ignores any clustering in the imputation model and a mixed effects imputation model. The study is restricted to a 2 level dataset with missing values observed in the continuous outcome variable in a balanced study design. However, cohort studies, multicentre studies and surveys often require clusters to be modelled as random effects (see section 3.1.1). This requires us to resort to multilevel analysis, where higher level units are modelled as random.

In their simulation study, Andridge evaluates the bias when covariates are available under MCAR and MAR and varies both cluster sizes and ICC values. Andridge (2011) concludes that imputation with fixed effects leads to over coverage for small ICC values and ignoring the clusters results in under coverage for large ICC values. Coverage is at or near nominal for random effects imputation model across all values of ICC. He demonstrates that multiple imputation of cluster randomized trials with an imputation model incorporating clusters as fixed effects, can result in severe overestimation of the variance of group means which is more severe in cases with small cluster sizes and small ICC values. Strong covariate information can help reduce bias in the MI variance estimator, although strong covariates are not always present in the dataset.

Andridge (2011) draws attention to several research papers (Browne et al., 2001; Clarke et al., 2010) that had to resort to specifying dummy variables to incorporate for the clustering in the dataset; suggesting an unawareness among researchers on the proper way to account for the data structure in the imputation model. Andridge acknowledges that a limitation in the study is the assumption that all observations were independent for all subjects while in reality, the opposite is true in clustered study designs. In this thesis, we are able to set aside this assumption.

Across literature, we see a shift towards development and creating a rich imputation model. (Yucel, 2008) and (Andridge, 2011) emphasizes the importance of a rich imputation model. With MI becoming a popular approach to handling missing val-

ues, the phrase ***"rich imputation model"*** has more significance than before.

3.4.1 Diagnostic Tools

There remains a limitation in the diagnostic tools available to check the appropriateness of an imputation model. Abayomi, Gelman, and Levy (2008) developed a series of internal and external checks to help diagnose the validity of the imputation models. Several other studies (Bondarenko & Raghunathan, 2016; Nguyen et al., 2017a) have addressed this gap by providing guidance on performing imputation diagnostics, such as exploring the imputation values through graphical displays like histograms and boxplots. These are referred to as external checks as the model is being assessed using information that is external to the data available.

Other diagnostic checks include comparison of the imputed and observed data. This is one of the most popular methods of diagnostic check and is referred to as an **internal check** since the imputed data is being assessed with respect to the data in hand (van Buuren, 2018). These checks help the user to identify potential issues with the imputation model. These are key checks to consider in this thesis and by extension in any multilevel MI analysis, as the most challenging task in MI is the specification of the imputation model. Any misspecification in the model would result in biased estimates. However, these methods are not perfect and have certain limitations. It can be difficult to perform these checks for every imputed values in a dataset with large variables. These tools cannot be used to test the validity of the MAR assumption without knowing the values of the missing data which in practice is not feasible (Nguyen, Carlin, & Lee, 2017b). However, it is beyond the scope of this thesis to develop these diagnostic tools to overcome their limitations. We use several of these tools to check the validity of the imputation models along with other graphical displays to confirm a consistency in the relationship between variables in both the actual and imputed datasets (See Chapters 5 and 6).

3.4.2 Handling Derived Variables

Another key aspect of related work that needs to be discussed is handling derived variables (e.g. cluster means and deviations) that are highly prevalent in medical and health research. A key paper by (von Hippel, 2009) (2009) focused on defining derived variables during imputation. This study recognises that any relationship in the analysis model should also be a part of the imputation model in order to ensure congeniality between the imputation and analysis model. This paper explores the different ways in which MI can account for this relationship. In his study, the relationships of interest are non-linear transformations of regressors such as squares, interactions, etc. In his study, von Hippel (2009) describes two methods *1) transform, then impute, 2) impute then transform*.

In addition, von Hippel demonstrates a third approach called passive imputation (see sections 2.3.9) which the author refers to as a more *disguised variant* of the impute then transform method. In Passive Imputation, the derived variables are used as regressors but are not imputed as dependent variables. The original variable is first imputed and the 'derived' variables are then recalculated using the imputed values.

In his paper, von Hippel emphasises that it is important that the imputed values reproduce the same mean and covariance matrix as the complete data. He advocates the use of the transform then impute method (or "Just Another Variable" (JAV)) when imputing transformed variables.

In his study, von Hippel departs from the original theme of the thesis that focuses on multilevel MI. He focuses purely on non-linear transformations and interactions terms, and has no extrapolations within the multilevel setting. However, I believe this paper is crucial as an important aspect in this thesis is centred around the inclusion of cluster means as derived and auxiliary variables in the imputation and

analysis models. It is important to understand the method will be using to handle these derived variables. This thesis heavily relies on passive imputation to recalculate the cluster means from the imputed variables at each iteration.

3.4.3 The role of technology in MI

Until a few years ago, the typical solution to handling missing values in Level 2 predictors was to delete all records in the cluster. Extending this solution to missing values in level 3 clusters, would not only result in dramatic reduction of sample size during analysis but is a terrible waste of resources such as manpower, money or time spent on collecting this information. We can positively note that technology has since improved allowing users to perform multiple imputation for missing data across software. Rubright, Nandakumar, and Gluttin (2014) compared multiple imputation across different proportions of missingness in the data. However, the current literature on missing data techniques for different proportions of missingness for complex data structures is limited. Multilevel data structures are commonly found in social science research (Goldstein & Rasbash, 1996; Chandola, Clarke, Wiggins, & Bartley, 2005; Subramanian, Kim, & Kawachi, 2002).

Enders et al. (2016) are probably among the first few to use **simulations**to address missing data in multilevel models. They compared the performance of four missing data techniques: single level imputation, single level imputation using dummy coded level 2 variable, JoMo and MICE. They concluded that MICE and JoMo are appropriate for random intercept models, with JoMo proving superior to MICE in analysis that included both level 1 predictors and its cluster mean. This is encouraging as the following chapters focus on addressing multiple imputation using MICE and JoMo in a random intercept model with cluster level variables that measure within and between cluster deviations. The paper focuses on missing values in the outcome variable and a covariate (X) that is measured at level 1. However, Enders' random

intercept models include only cluster means as predictor variables and do not focus on individual deviations in variables. We can speculate that as the number of levels increases in the data, using only cluster means; for say a variable (X_{ijk}) measured as the level 1; as a predictor may prove harmful and may result in inflation of standard errors.

Enders et al. (2016) acknowledge that the software hasn't developed enough to impute for missing values in Level 2 and **speculate that the Gelman and Hill method of separate imputations across levels would be ideal for imputing missing data in level 2 variables.** They deviate from the method proposed by Gelman and Hill by proposing a stepwise scheme where the researcher applies single level imputations to impute values at a higher level. After the imputations converge, the imputer uses the imputed level 2 variables as predictors for the level 1 imputed models. However, the paper fails to demonstrate this procedure in their analysis. Software has since then developed a great deal with imputations available for variables up to level 2. The study does not focus on the impact of varying cluster sizes on MI.

In his book, van Buuren (2018) has an entire chapter dedicated to the application of multilevel multiple imputation using R. He was the first to propose the use of MICE in the multilevel context. This book provides a key source of literature on the implementation and analysis of multilevel multiple imputation. In it, van Buuren reviews and demonstrates the current process in place for multilevel MI using the current packages MICE and JoMo in R. He demonstrates how to handle missing values in level 2 predictor variables. Like Gelman & Hill, and Yucel, van Buuren addresses the need to include cluster means as a predictor and confirms that adding cluster means in the imputation model is preferable.van Buuren explored the different methods of imputation namely MICE and JoMo in R. The book provides as a fantastic guide to the implementation of MI using R.

Practical applicability of MI

In the multilevel MI space, van Buuren (2018) addressed the occurrence of missing data in up to 2 levels under different scenarios such as a binary or continuous variables. This work is crucial as it displays the practical applicability of MI in R and serves as a guide on how to implement multilevel MI, albeit its applicability is limited to missing values up to level 2. The book release by van Buuren (2018) is a practical guide and checks the performance of methods MICE and JoMo in R for imputing missing data in two level models. van Buuren provides a recipe to impute missing values for level 1 and level 2 target variables, and demonstrated the application of this recipe using an example. Since the example dataset (used for demonstration) is a subset of the complete data and thus a subset of the possible analysis model, it is possible that several auxiliary variables may not have been modelled or accounted for. While this serves as a fantastic way to familiarise novice users with MI, it may not be directly replicable in a real world dataset.

In his demonstration, van Buuren restricted the number of variables in his dataset to six for simplicity in demonstration, and there is not enough information on the mechanisms driving the missing data in the example. There is some ambiguity in the number of units in level 1 and 2. van Buuren also calls for further study to be done to increase experience when handling imputation in datasets with three or more levels.

Lüdtke et al. (2017) conduct a simulation study to provide guidance on using MI in multilevel models with random intercept, random slope and cross level interactions. The two level random intercept model accounts for the clustered structure of the data by including random effects for each level 2 unit (Snijders & Bosker, 2011). Secondly, the study differentiates between the effects of level 1 covariates at individual and group level.

The study implements MICE and JoMo for analysis. Lüdtke et al. (2017) conclude that when missing values occur in level 1 predictor variables, both processes yield

approximately the same parameter estimates making both MICE and JoMo suitable for general applications in the multilevel random intercept model. Multilevel MI proves more challenging when a random slopes model is employed to handle missing values in level 1 covariates. While considering imputation for missing values at higher level variables, a random intercept model is developed with 25% MAR introduced in the level 2 covariate. The number of units at the second level were varied at 100 and 500. For missing values in level 2 covariate, imputations are generated by incorporating the group level information supplied by level 1 predictors and outcome variable. This is an idea that we have inherited in the thesis as it employs the Gelman and Hill method of separating imputations at each level and using aggregate measures of lower level predictor variables for imputation. We note that all the variables generated here were assumed to be standardized normal. The authors consider only one variable at each level in the model.

Lüdtke et al. (2017) identify a gap in literature assessing the performance of multilevel MI in more specialized situations including settings with very few units at the individual level, or the higher level, different mechanisms of missingness, and in models with higher levels of hierarchy. In principle, Lüdtke et al. theorize that MI would be possible with the existing software, however its implementation is yet to be seen.

This thesis address this gap by investigating the application of multilevel MI in three level models through simulations and real world applications. The thesis also addresses the inclusion of auxiliary variables in the multilevel imputation model. These are variables that are related to the mechanism of missingness in the data or are correlated to the variables with the missing values themselves (Dong & Peng, 2013). Missing values can be inferred from the observed data with more accuracy when more information can be included from the auxiliary variables (Schafer & Graham, 2002; Enders, 2008).

3.5 Goals of the thesis

Based on Figure 3.4.2 and in alignment with the research themes stated in Section 1.2, we identify the gaps in the literature on multilevel multiple imputation, and assess how to tackle them. Using this, we define specific objectives within the goals of the thesis as follows:

1. To identify methods of imputation for dealing with incomplete data for more than 2 levels and implement them using current statistical software.

2. To compare performance of extended MI methods (MICE & JoMo) under different types of missing (MCAR/MAR) and different proportions of missingness across different levels in 3 level data

   (a) Compare performance of extended MI methods for missingness occurring together across all levels in the dataset.

   (b) Compare the performance of both methods for handling derived variables in clustered datasets.

   (c) To measure the impact of the proportion of missing values at each level when developing the imputation model.

3. To apply these methods to simulated and real datasets.

3.6 Summary

In this chapter, we introduced our three-level hierarchical model, that will be the foundation model on which our research is based. The chapter also defined recurring concepts like fixed and random effects, and centering of covariates around the group mean. Finally the chapter gave an overview of the literature and body of work in the intersection of multiple imputation of multilevel models and identified the gap in literature within which this thesis rests. We defined the goals of the thesis using this information. The next chapter will describe how we engineered multiple imputation for three level datasets, the various combination of packages in R that

allowed us to achieve this, a detailed description of the simulation study set-up, and the mechanism that was used to introduce missingness in variables across each level in the dataset. The chapter also extends van Buuren procedure for imputing higher level variables to level 3 or more variables with missing values. The next chapter also provides a detailed description of data generation for the simulation study.

Chapter 4

Extension and Implementation of Multiple Imputation to Three Level Models

The purpose of this chapter is to illustrate the use of techniques of multilevel multiple imputation, and demonstrate the use of visualization tools to investigate the degree of missingness in the dataset as well as to study the performance of the imputation methods in three level models.
This chapter helps break down the structure and procedures of multilevel multiple imputation using MICE and JoMo in R (R Core Team, 2018) and describes the combination of R tools that allow us to perform this successfully. The results of the simulation described here are reported in the following chapters, while this chapter serves as a practical guide for those intending to use MI in practice for three level data structures.

Missingness is not usually the main focus of the study and dealing with it in a principled manner is often both computationally and conceptually challenging. Researchers often end up ignoring these missing values or resorting to unconventional methods with the intention of *completing* the datasets. To reduce complexity in analysis and to demonstrate how to perform MI in R, we have generated missing values that are at random (MAR). This chapter will serve as a scaffolding for the analysis performed in the simulation study in the following chapter. We will perform Com-

plete Case Analysis along with MICE and JoMo and compare the estimates obtained from the three methods.

4.1 Tools required to extend Multilevel Multiple Imputation to three level models using R

Multiple Imputation (MI) gets trickier when dealing with multilevel data. Despite the increase in awareness about the problems around imputation of missing values in multilevel data, there is limited literature on how to account for these features in Multilevel MI. While, multilevel multiple imputation is gaining recognition across statistical software, there is still some ambiguity on how to perform multilevel multiple imputation in practice. Recent years have seen the introduction of several packages in R (R Core Team, 2018) for this purpose. In this section we explore popular packages that have been used for multilevel MI in R. These can be understood as tools or ingredients used to follow the recipe described by van Buuren (see Table 4.1.1). Visualisation of the dataset can be used as an effective tool to understand the degree or extent of missingness in the dataset (Templ & Filzmoser, 2008). It can also help throw light on possible factors driving missingness especially in large datasets. This section commences with an introduction to the several packages and tools that can be used for visualization, and then goes on to the tools for imputation using R.

Table 4.1.1: Algorithm for imputing missing values in level 1 & level 2 target variables

---

Algorithm for imputing missing values in level 1 target variable

1. Define the most general analytic model to be applied to imputed data.
2. The target variable is the variable with the missing values.
3. Select an imputation method according to the target variable distribution.
4. Include all level 1 variables and their cluster means.
5. Include all level 2 predictors.
6. Include any interactions implied by the model.
7. Exclude any terms involving the target variable.
8. Implement the selected method and check if it imputes close to the observed data.

Algorithm for imputing missing values in Level 2 target variable

1. Define the most general analytic model to be applied to imputed data.
2. The target variable is the variable with the missing values.
3. S Select an imputation method according to the target variable distribution.
4. Include cluster means of all level 1 variables and interactions.
5. Include all level 2 predictors and interactions.
6. Exclude any terms involving the target variable.
7. Implement the selected method and check if it imputes close to the observed data.

Recipe adapted from (van Buuren, 2018).

4.1.1 `VIM` & `ggplot2`

The package **`VIM`** or "Visualization and Imputation of Missing Data" was released in October 2011 (Templ, Alfons, Kowarik, & Prantner, 2011). It provides new tools for visualising both missing and imputed values. This can be used to help identify possible trends in missing values, and the mechanisms generating missing data. `VIM` cannot be used as a diagnostic tool to prove that the missing data is MCAR or MAR, but can be used to provide initial insight into missingess in the dataset. It should be first used to look at the dataset containing missing values and then again after imputation to help compare the missing and the imputed values in the dataset. There are four single imputation techniques implemented in `VIM`, namely, hot deck imputation, k-NN imputation, regression imputation and robust model based imputation. Further details of imputation using `VIM` is discussed in the paper by Templ and Kowarik,(Kowarik & Templ, 2016).

While **`ggplot2`** is a package used broadly for data visualization in R, it has several functions specifically for exploration of missing data mechanisms. ggplot2 allows exploration of different missing data mechanisms, b incorporating of shifting of missing values so that we can visualise them on the same axes as the observed data. `geom_miss_point` is a function that can be applied to transform and plot missing values in ggplot2.

4.1.2 `mice`, `miceadds` & `micemd`

The package "Multiple Imputation by Chained Equations" or `mice` was developed by van Buuren in 2010 (van Buuren et al., 2015). This package provides multiple imputation using the chained equations method (see section 2.3.2). It allows the user to not only impute the data but also generate multivariate missing values for simulation purposes. Users can graphically compare the distribution of the imputed versus observed plot. Analysis can be performed on each of the imputed datasets separately and pooling of the results is also simplified using this package. **`mice`** allows the user to specify the number of imputations, the method of imputation to be used and the

number of iterations of the Gibbs Sampler.

The package **micemd** or "Multiple Imputation by Chained Equations with Multilevel Data" is an add-on package to the `mice` and is used to perform MI using FCS in datasets with 2 levels (Audigier & Resche-Rigon, 2017). It was released by Vincent Audigier on 23rd August 2017. It includes imputation methods for sporadic and systematically missing values, and allows the user to decide the method of imputation to be used based on the number of clusters, cluster size and type of variable (continuous, binary or integer).

The package **miceadds** includes "Additional MI functions for `mice`" (Robitzsch, Grund, Henke, & Robitzsch, 2019). This recent feature is an exciting addition to the package as it is based on the linear mixed effects model implemented in the **lmer** function allowing the imputation model to have multivariate random effects.

4.1.3 jomo

The package **jomo** or "Multilevel Joint Modelling Multiple Imputation" was released by Matteo Quartagno and James Carpenter in 2017 (Quartagno & Carpenter, 2016). This package incorporates multiple imputation using joint modelling (see Section 2.3.4). It allows the user to specify level 1 and level 2 target variables in the imputation models separately along with the data frame and the predictor variables in the multivariate joint imputation model (see the model described in section 2.3.4) . `jomo`can easily deal with interactions and polynomial terms in the model. It specifies a congenial imputation model when the model of interest in a generalized linear mixed effect model. Like `mice`, `jomo` allows the user to compare the distributions of imputed and observed values, and check the convergence of the imputation model.

4.1.4 lme4

While unrelated to multiple imputation, we applied the function **lmer** within the package **lme4** (Pinheiro & Bates, 2000) to fit the linear mixed models which were

used for fitting both the analysis and the imputation model. The function allows us to incorporate both fixed effect and random effects to run a mixed-effects model, which is pertinent to our study. With our multilevel model with three level, we require a multiple random effects expression to accommodate nesting of level 1 units within level 2, and level 2 units within level 3 units. The `lmer` function is designed to accommodate multiple nesting and correct model specifications.

4.2 Simulation Study Setup

To illustrate the use of these packages for three level multiple imputation using R, we use a 3 level simulated dataset. The structure of this simulation set-up mimics that of the Demographic Health Survey data structure, where each individual is nested within a household which is further nested within each Primary Sampling Unit (PSU).

In India, the DHS survey is referred to as the National Family Health Survey (NFHS), the most recent being the NFHS- 5 (2019-2020) survey. However, this data was not released during the duration of writing this thesis, and the fourth NFHS survey was used for analysis. The fourth National Family Health Survey (NFHS-4) is a nationally representative household survey coordinated by the International Institute for Population Sciences (IIPS) under the aegis of the Government of India, was conducted in 2015-2016. As did NFHS-1 (1992-93) and NFHS-2 (1998-99), NFHS-3 (2005-2006); NFHS-4 also provides population based estimates on fertility, mortality, family planning, HIV-related knowledge, and important aspects of nutrition, health, and health care for women aged between 15-49 years and men aged between 15-54 years.

Different information is captured at each level (i.e. individual, household and PSU) and is further aggregated to the higher level unit. We include cluster means in the imputation model as it proves advantageous in all circumstances containing multilevel structures (Enders, 2008; Mistler & Enders, 2017; van Buuren, 2018). Adding these *aggregates* or *contextual variables* allows us to estimate both within and between

group regressions. Cluster means are valid predictors for large and equal cluster sizes. The random intercept model below is the analysis model used in the simulation study setup where Y_{ijk} is the continuous measure for the i^{th} individual in the j^{th} household in the k^{th} PSU (Menon et al., 2019).

We recall that in this thesis, we identify variables with the subscripts i,j and k to signal to the level at which the variable is captured. For example, X_{ijk} is a variable captured at the lowest level (Level 1) where (i) refers to the number of units at the lowest level, (j) refers to the second level (Level 2) and (k) refers to the third level (Level 3). Here, the subscript (i) denotes the individuals, (j) denotes the households and (k) denotes the districts. Readers should note that throughout this thesis, the covariates measured at the lowest level are signalled to by the letter X, the covariates at the second level are denoted by the letter Z and the covariates at the highest levels are denoted by the letter W. Outcome variables are denoted by the letter Y. So in this thesis, X_{ijk} denotes that the variable was captured for the i^{th} individual in the j^{th} household in the k^{th} PSU; and Z_{jk} is the variable was captured for the j^{th} household in the k^{th} PSU.

Average family size in India is approximately four individuals per household. In order to reflect the same, we set number of individuals/household = 4, number of households/PSU = 20 and number of PSUs = 100. This was done to define a ratio of 1:5:25 between the number of units at each level. This choice was made to maintain the observed structure of the NFHS survey data. The model for the simulation study is therefore:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}(X_{ijk} - \bar{X}_{jk}) + e_{ijk};\ e_{ijk} \sim N(0, \sigma^2)$$

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01}(Z_{jk} - \bar{Z}_k) + \gamma_{02}(\bar{X}_{jk} - \bar{X}_k) + r_{0jk};\ r_{0jk} \sim N(0, \tau_r)$$

$$\beta_{1jk} = \gamma_{10}$$

$$\gamma_{00k} = \delta_{000} + \delta_{001}(W_k) + u_{00k};\ u_{00k} \sim N(0, \tau_u)$$

Substituting, we get

$$Y_{ijk} = \delta_{000} + \gamma_{10}(X_{ijk} - \bar{X}_{jk}) + \gamma_{02}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{01}(Z_{jk} - \bar{Z}_k) + \delta_{001}W_k + u_{00k} + r_{0jk} + e_{ijk} \tag{4.1}$$

To retain the data structure specified above i.e. number of individuals/household = 4, number of households/PSU = 20 and number of PSUs = 100, we define $n_{jk} = 4, n_k = 20$ and $K = 100$.
The parameter values for model (4.1) are provided in Table 4.2.3 and Table 4.2.2.

The model (4.1) explains how the total variation in the outcome variable is allocated across the three levels i.e. individuals within households(σ^2), households within districts (τ_r) and among districts (τ_u) (Menon et al., 2019). We also note that the the variables at level 1 (X_{ijk}) and level 2 (Z_{jk}) are centred around their group means. The method used to generate the predictor and outcome variables in the simulated dataset is explained in section 4.2.1. Unlike the NFHS study, we have simulated a balanced data structure, with fixed number of units within each higher level units. We also acknowledge that most statistical illustrations assume that the data follows a Normal distribution, however this is almost never true. We accommodate for non-normal distributions ($X_{ijk} \sim SN(\epsilon = 70, \omega = 20, \alpha = 10)$ in our analysis.

In this study, a random intercept model was fit with Y_{ijk} as the outcome variable and X_{ijk}, $X_{ijk} - \bar{X}_{jk}$, $\bar{X}_{jk} - \bar{X}_k$, $Z_{jk} - \bar{Z}_k$ and W_k as predictor variables. The dataset is hierarchical with 3 levels – "individuals" at level 1, "households" at level 2 and "PSUs" at level 3. Plotting Y_{ijk} against the predictor variables, we see a significant positive correlation has been generated between the variables Y_{ijk} and X_{ijk} (r=0.59, p $<$ 0.001), Z_{jk}(r= 0.72, $p < 0.001$) and W_k (r = 0.22, $p < 0.001$). In Figure 4.2.1, we can see the distribution and relationship between the variables in the dataset. The positive correlation between X_{ijk} and Y_{ijk} is observed in the top left plot. We also observe the distribution of variables X_{ijk}, Y_{ijk} across various levels of W_k which is a count variable of mean value = 3, and the distribution of X_{ijk}, Y_{ijk} across various levels of W_k which is a count variable of mean value = 3.

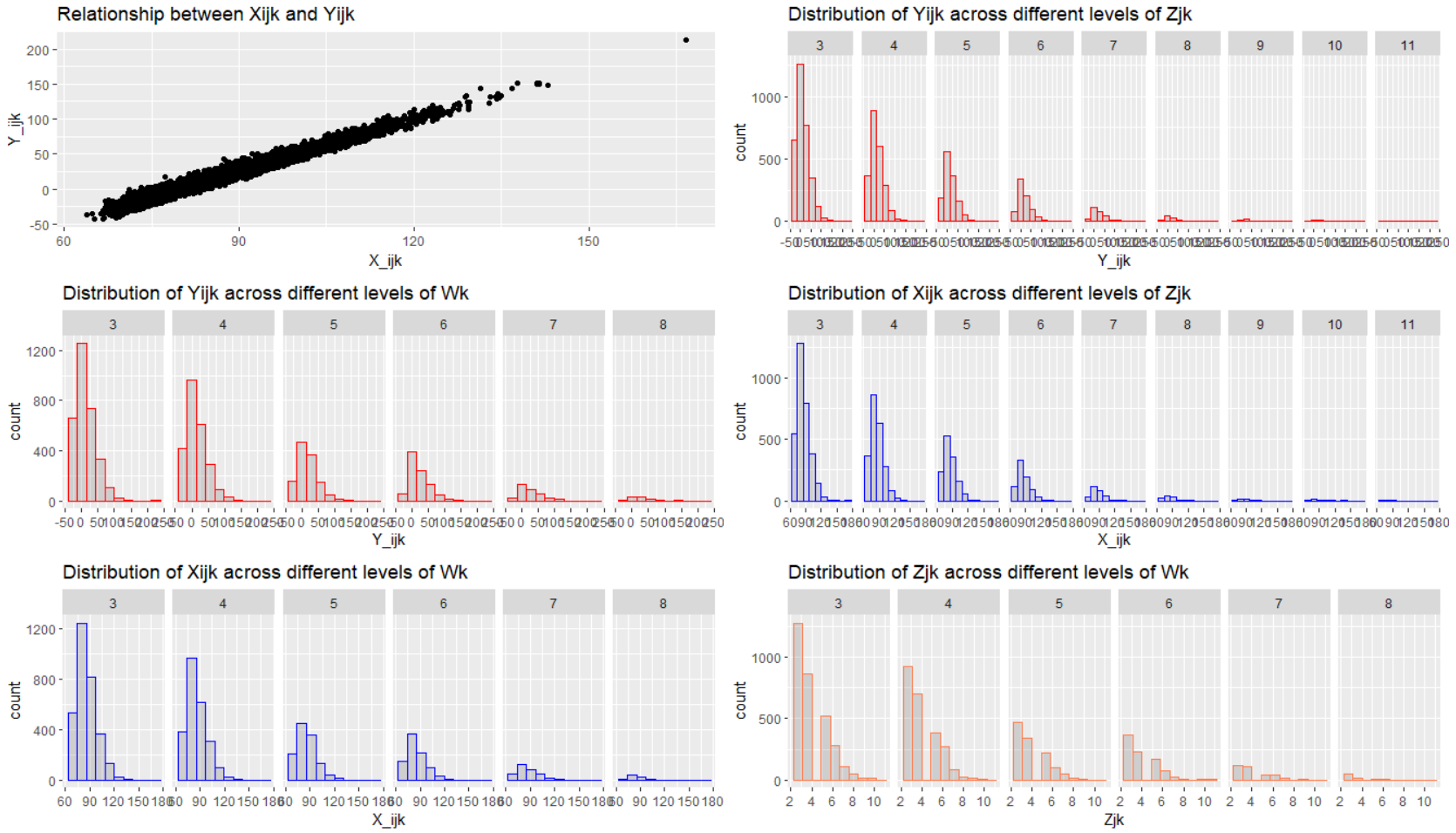


Figure 4.2.1: Relationship between variables in the dataset

Random intercept models are often referred to variance component models as the residuals are segregated into components corresponding to each level in the dataset. Here u_{00k} represents the deviation of the j^{th} household from the average prediction line. Hence, the vectors r_{0jk} contains entries for the level 2 residuals and contains entries for level 1 residuals. While r_{0jk} and u_{00k} modifies the intercept term, the slope coefficient is fixed. Centering variables in multilevel modelling have shown to be helpful in terms of providing stability and computational ease (Gelman & Hill, 2006; Kreft & De Leeuw, 1998). Centering can also be useful in separating individual level effects from group level effects.

The package `lme4` (see section 4.1.4) is used to perform mixed effects linear regression in *R*. A mixed effect model could be either a random intercept model, a random slope model or both. We have three variants that we generally consider when defining a mixed model. These are defined below and the corresponding code; so as to accommodate these variations in function `lme4`; is specified in the corresponding brackets.

- Intercept only by random factor : $(1|randomfactor)$
- Slope only by random factor : $(0 + fixedfactor|randomfactor)$
- Intercept and Slope by random factor : $(1 + fixedfactor|randomfactor)$

If we assume that the intercept and slope are calculated independently i.e. without any assumed correlation, then we define a 4th variant

- Intercept and slope separated by random factor:
  1. $(1|randomfactor) + (0 + fixedfactor|randomfactor)$
  2. $(fixedfactor) + (fixedfactor|randomfactor)$

4.2.1 Data generating mechanisms

We generated 1000 datasets of 8000 observations each (4^*20^*100), i.e. 4 units at level 1, 20 units at level 2 and 100 units at level 3. Level 1 and Level 2 variables are centered around their cluster means. Although imputations are reasonably faster with

the advancements in technology, beyond 1000 the gains in terms of understanding the variability of imputation estimates is small. Across literature, we see a range of 500 to 10,000 simulations, so the value chosen for this study was well within this limit. The data generating mechanisms were inspired by routinely captured variables in the NFHS surveys. X_{ijk} was inspired by variables such as Systolic blood pressure values for the i^{th} individual in the j^{th} household in the k^{th} PSU. X_{2ijk} which is highly correlated to the variable X_{ijk} was generated as Diastolic blood pressure values for the i^{th} individual in the j^{th} household in the k^{th} PSU. Variable Z_{jk} and Z_{2jk} are the number of children, and the number of bedrooms in each household respectively. Finally, variable Wk is the total number of Anganwadis in the k^{th} PSU. The analysis model is given in equation 4.1. We set the following parameters as: $\delta_{000} = 2, \gamma_{10} = 2.5, \gamma_{02} = 2.5, \gamma_{01} = 2$, and $\delta_{001} = 3$. These are the true values against which we compare all our estimates obtained using various missing data analysis methods. We fitted a multilevel model on the first simulated dataset, as dry run (without any missing values) to check if we can retrieve the true values of the parameter estimates. The results of this "dry run" are presented in Table 4.2.3. This gives us a rough approximation of what our results should be after we analyse using the missing data methods. Centering variables in multilevel modelling have shown to be helpful in terms of providing stability and computational ease (Gelman & Hill, 2006; Kreft & De Leeuw, 1998). Centering can also be useful in separating individual level effects from group level effects. We also included variables (X_{2ijk}) ans (Z_{2jk}) that are positively and strongly correlated with variables (X_{ijk}) and (Z_{jk}) respectively. Corr$(Z_{jk}, Z_{2jk}) \approx 0.8$, Corr$(X_{ijk}, X_{2ijk}) \approx 0.8$. We simulate X_{ijk} that are MAR condition of X_{2ijk} and Y_{ijk}, Z_{jk} that are MAR conditional on Z_{2jk} and $\bar{Y_{jk}}$ and W_k that are MAR on $\bar{Y_k}$.

We simulate the dataset to mimic a multilevel household survey with individuals nested within households further nested within primary sampling units (PSUs). Thus the variable X_{ijk} represents the information captured for the i^{th} individual in the j^{th} household within the k^{th} PSU; $(i = 1, 2...4); j = (1, 2, ..20)$ and $(k = 1, 2...100)$. For ease in calculation and analysis, we set the clusters (households and PSUs) of

Table 4.2.2: Variable Generation in Simulation Study

Variable	*Variable Generating Procedure*
Level 3	Number of higher level units e.g. PSUs $(k = 100)$
Level 2 (middle)	Number of higher level units each nested within level 3 units e.g. households $(j = 20)$
Level 1	Number of lower level units each nested within level 2 units, further nested within level 3 units e.g. individuals $(i = 4)$
$e_{ijk} \sim N(0,1)$	Level 1 error term
$r_{0jk} \sim N(0,1)$	Level 2 error term
$u_{00k} \sim N(0,1)$	Level 3 error term
Coefficients	$g_{000} = 2,\ g_{100} = 2.5,\ g_{200} = 2.5,\ g_{010} = 2,\ g_{001} = 3$
Y_{ijk}	Level 1 outcome variable (continuous)
$X_{ijk} \sim SN(\varepsilon = 70, \omega = 20, \alpha = 10)$	Level 1 predictor variable (skewed normal (SN))
$Z_{jk} \sim Pois(\lambda = 3)$	Level 2 predictor variable
$W_k \sim Pois(\lambda = 3)$	Level 3 predictor variable
Correlated Variables:	
X_{2ijk}	Level 1 predictor variable (Corr(X_{ijk}, X_{2ijk}) = 0.8)
Z_{2jk}	Level 2 predictor variable (Corr(Z_{jk}, Z_{2jk}) = 0.8)
Derived Variables:	
$\bar{X}_{jk}$	Mean of (X_{ijk}) calculated for each level 2 unit
$\bar{Z}_k$	Mean of (Z_{jk}) calculated for each level 3 unit
$\bar{X}_k$	Mean of (X_{ijk}) calculated for each level 3 unit
$X_{ijk} - \bar{X}_{jk}$	Deviation of each level 1 observation (X_{ijk}) from the level 2 mean
$\bar{X}_{jk} - \bar{X}_k$	Deviation of each level 2 mean from the level 3 mean
$Z_{jk} - \bar{Z}_k$	Deviation of each level 2 observation (Z_{jk}) from the level 3 mean

Table 4.2.3: Estimates obtained from running the multilevel model on the first simulated dataset

Predictors	*True Parameter*	*Estimates*	*95% CI*
(Intercept)	$\delta_{000} = 2$	2.64	1.93 - 3.34
$X_{ijk} - \bar{X}_{jk}$	$\gamma_{10} = 2.5$	2.50	2.50 - 2.52
$\bar{X}_{jk} - \bar{X}_{k}$	$\gamma_{02} = 2.5$	2.50	2.49 - 2.51
$Z_{jk} - \bar{Z}_{k}$	$\gamma_{01} = 2$	2.00	1.96 - 2.03
W_k	$\delta_{001} = 3$	2.87	2.70 - 3.04
σ^2		1.01	
$\tau_{00\ level2:level3}$		1.06	
$\tau_{00\ level3}$		1.00	
$ICC_{level2:level3}$		0.35	
ICC_{level3}		0.33	
Observations	$n = 8000$		

equal size. The analysis model is given as:

$$Y_{ijk} = \beta_{0jk} + \delta_{00k} W_k + e_{ijk}; (i = 1, 2, ...n; j = 1, 2...J, \& k = 1, 2...K)$$

$$\beta_{0jk} = \gamma_{00k} + \gamma_{10}(X_{ijk} - \bar{X}_{jk}) + r_{0jk}$$

$$\gamma_{00k} = \delta_{000} + \gamma_{02}(X_{jk} - \bar{X}_{k}) + \gamma_{01}(Z_{jk} - \bar{Z}_{k}) + u_{00k}$$

Missing values were introduced into the dataset under the MAR mechanism, where the probability of missingness was determined by a logistic regression model (Lee & Carlin, 2017). Data generation for X_{2ijk} and Z_{2jk} are explained in table 4.2.2. Each unit's binary response variable is drawn from a Bernoulli distribution:
Here R_{ijk} is **Probability of MAR in X_{ijk} was determined by model**

$$P(R_{ijk} = 1) = \frac{e^{X_2^{(s)} + \beta Y^{(s)}}}{1 + e^{X_2^{(s)} + \beta Y^{(s)}}};\ where\ Y^{(s)} = \frac{Y - E(Y)}{SD(Y)} \& X_2^{(s)} = \frac{X_{2ijk} - E(X_{2ijk})}{SD(X_{2ijk})}$$

Probability of MAR in Z_{jk} was determined by model

$$P(R_{jk}=1)=\frac{e^{Z_2^{(s)}+\beta Y'^{(s)}}}{1+e^{Z_2^{(s)}+\beta Y'^{(s)}}};\ where\ Y'^{(s)}=\frac{\bar{Y}_{jk}-E(\bar{Y}_{jk})}{SD(\bar{Y}_{jk})}\&\ Z_2^{(s)}=\frac{Z_{2jk}-E(Z_{2jk})}{SD(Z_{2jk})}$$

Probability of MAR in W_k was determined by model

$$P(R_k=1)=\frac{e^{2+\beta Y^{*(s)}}}{1+e^{2+\beta Y^{*(s)}}};\ where\ Y^{*(s)}=\frac{\bar{Y}_k-E(\bar{Y}_k)}{SD(\bar{Y}_k)}$$

Here, β was set to -0.85 when generating 20% MAR and β was set to 0.85 when generating 50% MAR. Data generation of variables $X_{2,ijk}$ and $Z_{2,jk}$ is explained in the table 4.2.2. $X_{2,ijk}$ and $Z_{2,jk}$ are introduced as variables with complete information, that can be used to predict missing values in covariates captured as Level 1 and Level 2, to simulate the "Missing at Random" mechanisms in the the latter. Inclusion of these complete variables in the imputation models for X_{ijk} and Z_{jk} allows us to correctly impute under the assumption of MAR. It is important to note, that exclusion of these two variables in their corresponding imputation models would result in data in X_{ijk} and Z_{jk} being Missing Not at Random (MNAR). This however goes beyond the scope of this thesis, and will need to be further explored in future research.

Since the variables below are inspired by the demographic health survey data, the distribution of the variables were chosen to match the distribution of commonly observed biomarkers and other household level characteristics. The data generating mechanism is reflective of this property (see X_{ijk} in table 4.2.2).

4.2.2 Visualization and analysis of missing data in the first simulated dataset

Visualising missing data and identifying its presence in the study is the first step prior to analysing or imputing the results with missing values. We confirm that the dataset contains around 20% missingness in $X_{ijk}(24\%)$, $Z_{jk}(19\%)$ and no missingness in W_k. Visualizing the extent of missing values in the dataset, we first identify the different patterns of missingness observed in the dataset. This can be done using the

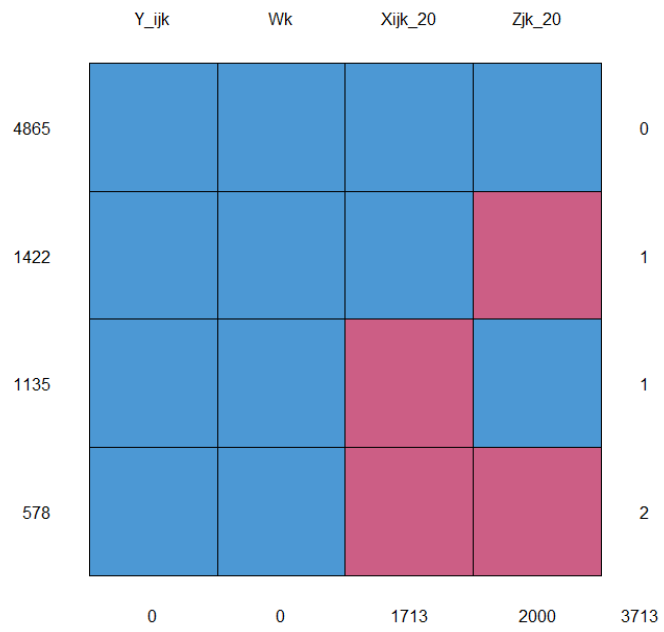


Figure 4.2.2: Patterns of missingness in the dataset

`md.pattern` command in `mice` (see Figure 4.2.2). There are four missing patterns observed in the dataset. Figure 4.2.2 can be interpreted as follows. We see that there are 4,865 cases with complete information across all variables. There are 1,422 observations with missing data in variable Z_{jk}, 1,135 observations with missing values in X_{ijk}, and 578 observations with missing values in Z_{jk} and X_{ijk}. There are no missing values in W_k in this dataset.

4.3 Specifying the Imputation Model

Performing Multiple Imputation using the wrong model can be more dangerous than Complete Case Analysis (van Buuren, 2007). Thus, the first step before we proceed to imputation of missing values, is to define the imputation model. It has been established that the analysis and imputation model should be congenial (see section 2.3.8). We need to be aware that the analysis model has derived variables as part of the analysis. These variables are cluster means of the level 1 predictor variable ($\bar{X}_{jk}$ and $\bar{X}_k$) or level 2 predictor variables ($\bar{Z}_k$). We can calculate these variables from the original variables, however this is possible only if the dataset is complete. There is no need to impute these values as they can be recalculated from the imputed (completed) datasets (van Buuren, 2018). However, we need to be aware of the role these variables play in the analysis model.

4.4 Multilevel Multiple Imputation: Engineering MICE (ext.)

In this method we repeat univariate multilevel imputation across multiple levels using MICE (van Buuren et al., 2015). This is done by drawing imputation by estimating parameters in every iteration.

Recall from section 2.3.2, the generalisation of MICE algorithms to iteration (t) and variable (j),

$$\theta_j^{(t)} \sim f(\theta_j).f(x_j^{obs(t-1)}|x - j^{obs(t)}\ y.\ \theta_j); (j = 1, 2, ..p)$$

.

$$x_j^{mis(t)} \sim f(x_j^{mis} | \ x - j^{obs(t)-1}. \ y. \ \theta_j^{(t)}); (j = 1, 2, ..p)$$

.

We alternate between each iteration and use the values of the imputed variable in the first step to complete the missing data in the second variable. The default number of iterations (t_{max}) = 5.

(Enders et al., 2016) have demonstrated that including the cluster means in the imputation model reduces bias of estimates and improved probability coverage of confidence intervals. This would imply that cluster means need to be calculated constantly from the imputed predictor variable in the previous iteration. In order to illustrate this effect, we include group mean- centred predictors for level 1 variables in the imputation model given below.

The imputation model to impute for missing values in X_{ijk} is specified below. Predictors in the imputation model for variable X_{ijk} include covariates at level 1 i.e. Y_{ijk} and $X_{2,ijk}$, covariates measured at level 2 i.e. $\bar{Y}_{jk}$ and $\bar{X}_{2,ijk}$, and covariates at the level 3, $\bar{Z}_k, W_k$ and $\bar{Y}_k$.

$$X_{ijk} \sim (\beta_0^* + \beta_1^*(Y_{ijk}) + \beta_2^*(\bar{Y}_{jk}) + \beta_3^*(X_{2,ijk}) + \beta_4^*(\bar{X}_{2,jk}) + \beta_5^*(\bar{Z}_k) + \beta_6^* W_k + \beta_7^* \bar{Y}_k + u_{00k} + r_{0jk}, e_{ijk})$$

Similarly, the imputation model to impute for missing values in Z_{jk} is specified below. Covariates measured at level 2 i.e. $Z_{2,jk}, \bar{Y}_{jk}$ and $\bar{X}_{2,jk}$, and covariates at the level 3, $\bar{Z}_k, W_k, \bar{X}_k$ and $\bar{Y}_k$.

$$Z_{jk} \sim (\gamma_{00}^* + \gamma_1^*(Z_{2,jk}) + \gamma_3^*(\bar{X}_k) + \gamma_4^*(\bar{Z}_{2k}) + \gamma_5^* \bar{Y}_k + u_{00k}, r_{0jk})$$

Finally, the imputation model to impute for missing values in W_k is specified below. Covariates at the level 3 are included as predictors in the model i.e. $\bar{Z}_k, \bar{X}_k$ and $\bar{Y}_k$.

$$W_k \sim (\delta_{000}^* + \delta_1^* \bar{Y}_k + \delta_2^*(\bar{X}_k) + \delta_3^*(\bar{Z}_k), u_{00k})$$

We use the *miceadds* package which contains functions to impute for missing values in multilevel models. It is also reasonable to estimate the cluster means when all X_{ijk} have been imputed and include them in the final analysis model. We note that the coefficients in the imputation model are not the same as the coefficients in the analysis model (refer Section 2.2.1). Looking at the imputation model for X_{ijk} , we see that it follows the recipe for imputation provided by van Buuren (2018), see Table 4.1.1. The imputation model contains the level 1 variables (including the outcome variable Y_{ijk}), cluster means of level 1 variables ($\bar{Y}_k$ & $\bar{Y}_{jk}$).We also note that the imputation model for X_{ijk} contains several variables such as (Y_{ijk}, $\bar{Z}_{jk}$) that are included in the analysis model and variables not included in the analysis model (such as $X_{2,ijk}$ and $Z_{2,jk}$). We can confirm that the two models are congenial with each other (Refer section 2.3.8). We now list the different components in MICE that needs to be defined before we proceed with imputation using MICE. This can be tied back to the elements described in Section 2.3.5.

Components to define within `mice ()` and `mice ()`

1. **Data frame:** We define a data frame with all the variables in the analysis model. This includes a column of 1_s to define the constant term.

2. **Imputation Method:** This is probably the most important component within mice. Choosing the right method of imputation is crucial when handling missing values. The method of imputation used is `ml.lmer` a function within `lmer` when used in combination with `miceadds`, which as the initials suggests is multilevel linear mixed effects modelling, and allows for imputation of variables in multiple levels in the dataset.The predictors are selected form the predictor matrix defined in `mice`.

3. **Predictor Matrix:** This matrix defines the variables that mice will use for imputation. By default, `mice` includes all variables in the dataset, but the user can redefine the matrix as per their requirement. This is where we define the imputation models for each target variable as specified above.

4. **maxit:** This is the number of iterations. Default value is 5.

5. **m:** This is the number of imputed datasets. Default value is 5.

The `variables_levels` function in `mice` allows us to specify the level at which the imputation occurs. This function allows the user to define the location of each variable in the model. The variables at the lowest level are left blank. Then, we define the predictors to be used for imputation of variables X_{ijk} and Z_{jk} with 20% missing values each.

4.4.1 How did we extend MICE to impute for missing values in the third level?

- We first used the command `make.predictorMatrix()` in `mice` to generate a valid predictor matrix. This is equivalent to the predictor matrix that would be generated by mice in a dry run. A dry run here refers to a `mice` run with no user - interference.

- We use this to then define user-specific setting within the predictor matrix, specifying predictor variables for the imputation model for X_{ijk} and Z_{jk}. After defining the predictor matrix, we still have to decide the method of imputation. As in the predictor matrix, we first use the command `make.method()`. We use the `ml.lmer` method to impute for missing values in Z_{jk}. This is done in R using the following chunk of code. The full code used for the simulations to impute for MCAR data using MICE is provided in Appendix A. The code shows specification of cluster membership of each variable, use of passive imputation to perform post imputation centering of variables, by careful modification of the existing `mice` package, used in combination with the package `miceadds` and `mitml`.

- We now define the hierarchical structure of the imputation model and specify the nesting observed in the data for each variable containing missing values. This is done using the `variables_levels` function in `mice` to specify the level

of each target variable and the covariates in the imputation model. For example, variables $X_{ijk}, Y_{ijk}, X_{2,ijk}, (X_{ijk} - \bar{X}_{jk})$ are set at level 1. Variables such as $\bar{X}_{jk}, Z_{jk}, Z_{2,jk}, \bar{Y}_{jk}, (\bar{X}_{jk} - \bar{X}_k), (\bar{Z}_{jk} - \bar{Z}_k)$, etc are set to level 2, and variables such as W_k, Y_k, X_k are set to level 3.

- We now use respecify `make.predictorMatrix()` function in `mice`, to specify the imputation models for each target variable as per the model defined above. We differentiate between the variables captured at same level by the value "3", and variables at the lower level by the value "1". The target variable at the lowest level is imputed first.

- Additionally, we use the cluster function in `mice` to specify nesting. For example, $cluster[[X_{ijk}]] = c("level2", "level3")$ and $cluster[[Z_{jk}]] = "level3"$. This implies that variable X_{ijk} is nested within level 2 further nested within level 3, and Z_{jk} is nested within level 3.

- Finally we specify the order in which `mice` imputed the data. We run the `mice` command and set maximum number of iterations = 10 and number of imputations = 5.

- Now that the variables with missing values have been imputed, we convert the `mice` object to a list, which allows us to use passive imputation to recalculate the derived variables using the recently imputed values. This is done to preserve the relations that existed between variables, before imputation was used. We reconvert it into a `mice` object to run the analysis model on the imputed values.

- Finally we run the analysis model using the `lme4::lmer` function within the `with` function which fits the analysis model for all 5 imputed datasets. The pooled estimates were obtained using the `pool`command which combine estimates using Rubin's rule.

This method extends and tests the methods specified in the blog written by Simon Grund (2019) on *Multiple imputation for three-level and cross-classified data*[1]. However in

[1]https://www.r-bloggers.com/multiple-imputation-for-three-level-and-cross-classified-data/

this study, we extended the method specified in the blog by extending Van Buuren's recipe (see Table 4.1.1) to level 3 or higher level target variables. This recipe has been used to help construct the imputation models used in both MICE and JoMo, especially in defining these within the software.

A key point to note is that unlike `mice`, `jomo` has not yet been developed to account for data structures with more than two levels. To accommodate for this and to ensure that two different comparisons were not being made between `mice` and `jomo`, we assume level 3 to be fixed in both algorithms. This approach enables us to impute for missing values in the third level using both methods.

Table 4.4.4: Algorithm for imputing missing values in level 3+ target variables

Algorithm for Level 3+ target variable

1. Define the most general analytic model to be applied to imputed data
2. The target variable is the variable with the missing values
3. Select a method that imputes close to the data
4. Include cluster means of lower level variables and interactions
5. Exclude any lower level variable that does not occur in the analysis model
6. Exclude any terms involving the target variable
7. Implement the selected method and check if it imputes close to the observed data.

Extrapolated from the recipe provided by van Buuren(2018). See Table 4.1.1.

4.5 Multilevel Multiple Imputation: Engineering JoMo (ext.)

In section2.3.4, we provided a deep insight into how JoMo works and defined a simple joint model and elaborated on the methods used to impute for missing values using JoMo. To recap, we defined a simple joint model given by the multivariate normal model as follows. Here, variables X_{1ij} and X_{2ij} are partially observed and

X_{3ij} andX_{4ij} are variables with no missing data.

$$X_{1,ij} = \beta_{01} + \beta_{11}X_{3,ij} + \beta_{21}X_{4,ij} + e_{1,ij} + u_{1,j}$$

$$X_{2,ij} = \beta_{02} + \beta_{12}X_{3,ij} + \beta_{22}X_{4,ij} + e_{2,ij} + u_{2,j}$$

$$\begin{pmatrix} e_{1,i} \\ e_{2,i} \end{pmatrix} \sim N(0, \Omega_e)$$

$$\begin{pmatrix} u_{1,j} \\ u_{2,j} \end{pmatrix} \sim N(0, \Omega_u)$$

In this section, we extend the joint modelling approach implemented in the package `jomo` (Quartagno & Carpenter, 2016) in R, to impute for missing values in the third level. In this method a single joint model is constructed for all variables with missing values (Lüdtke et al., 2017). The method employs a Gibbs sampler that allows both estimation of parameters as well as imputation of missing data in the joint imputation model (Carpenter & Kenward, 2012). The process has been explained in detail in section 2.3.4. As mentioned earlier, unlike MICE the software has not yet been developed to account for data structures with more than two levels. To accommodate for this limitation in software, I assume level 3 to be fixed in the model.

Components to define within JoMo ():

1. **Y –** a data frame with all the missing variables for the imputation model. JoMo assumes variables with missing values as outcomes for the imputation model.
2. **X –** a data frame with all the covariates for the imputation model.
3. **Z –** a data frame containing the covariates that are associated to the random effects in the imputation model.
4. **clus –** a data frame with the cluster indicators for each observation in the dataset.
5. **nburn –** Number of iterations. Default is 1000.

6. **nimp –** Number of imputations. Default is 5.

4.5.1 How did we extend JoMo to impute for missing values in the third level?

To impute for missing values upto the third level, we followed the procedure below:

- We defined the data frame (Y) which included all the target variables. This included any derived variables to be imputed. For example, to impute for missing values in X_{ijk}, Z_{jk} and W_k, we specify all variables including the derived variables that are $(X_{ijk} - \bar{X}_{jk})$, $(\bar{X}_{jk} - \bar{X}_k)$ and $(Z_{jk} - \bar{Z}_k)$ in the Y data frame.
- Next we specify all the covariates in the model in the dataset including variables such as $X_{2,ijk}$ and $Z_{2,jk}$.
- Next we specify the cluster identifier variable.
- We use the `jomo1ran` function within `jomo` to run the imputations.
- Now that the variables with missing values have been imputed, we convert the `jomomids` object to a list, which allows us to use passive imputation to recalculate the derived variables using the recently imputed values. This is done to preserve the relations that existed between variables, before imputation was used. We reconvert it into a `mids` object to run the analysis model on the imputed values.
- Cluster means for each variable was calculated using the `clusterMeans` function. For example, $\bar{Z}_k$ = clusterMeans$(Z_{jk}, level3)$.
- Finally we run the analysis model using the `lme4::lmer` function within the `with` function which fits the analysis model for all 5 imputed datasets. The pooled estimates were obtained using the `pool`command which combine estimates using Rubin's rule.

The imputation models are specified as follows:

$$X_{ijk} = \beta_{01} + \beta_{11}Y_{ijk} + \beta_{12}X_{2,ijk} + \beta_{13}Z_{2,jk} + \beta_{14}\bar{Y}_{jk} + \beta_{15}\bar{X}_{2jk} + \beta_{16}Z_{2,jk} + e_{ijk} + r_{jk} + u_{00k}$$

$$Z_{jk} = \beta_{02} + \beta_{21}Y_{ijk} + \beta_{22}X_{2,ijk} + \beta_{23}Z_{2,jk} + \beta_{24}\bar{Y}_{jk} + \beta_{25}\bar{X}_{2,jk} + \beta_{26}Z_{2,jk} + e_{ijk} + r_{jk} + u_{00k}$$

$$W_{k} = \beta_{03} + \beta_{31}Y_{ijk} + \beta_{32}X_{2,ijk} + \beta_{33}Z_{2,jk} + \beta_{34}\bar{Y}_{jk} + \beta_{35}\bar{X}_{2jk} + \beta_{36}Z_{2,jk} + e_{ijk} + r_{jk} + u_{00k}$$

The full code used for the simulations to impute for MCAR data using JoMo is provided in section 8.4 in Appendix A. We specify the cluster membership for each variable, use passive imputation to perform post imputation centering of variables, and carefully modify the existing `jomo` package which is used in combination with the package `miceadds` and `mitml`.

4.6 Summary

In this chapter, we have defined the underlying mechanisms of missingness that we use to introduce missing values in the dataset, and outlined the processes to extend both `mice` and `jomo`. We also describe the packages and tools required to impute for missing values in multilevel datasets. The simulation structure is described along with how to perform multiple imputation in three-level datasets. The next chapter describes the result of the simulations for a variety of scenarios and compares the performance of the methods of imputation using various measures.

Chapter 5

Simulation Study Part - 1

5.1 Comparing MICE (ext.) and JoMo (ext.) approaches

In the previous chapters, we acknowledged that use of multiple imputation when dealing with multilevel datasets is even more important than with datasets with independent observations. Missing observations can occur in variables captured at any or every level. They could occur at the lowest level or at the highest level or any combination of the two. They could occur in the outcome variable measure at the lowest level (individual), or predictor variables measured at level 1 (X_{ijk}) or level 2 (Z_{jk}) or level 3 (W_k), or missing values can be observed even at cluster identification.

Traditional methods such as Complete Case Analysis would not only remove observations from the lower level units, but presence of missing values in variables captured at higher levels, may result in deletion of all observations nested within that particular unit. For example, if we observe missing values for information captured at a household level, (e.g. number of members in the household), performing Complete Case Analysis would result in deletion of the information captured for all individuals in that household. Given the complexity of hierarchical data structures and the appropriate modelling procedures, multilevel datasets with missing values require special multilevel multiple imputation methods. Popular methods of imputation include MICE and JoMo (described previously). To further illustrate the ideas described in the previous chapters, we implement multilevel MICE (ext.) and JoMo (ext.) and compare their performance across various scenarios.

5.2 Aims

In this chapter, we compare performance of multilevel MI methods i.e. multilevel MICE (ext.) (see section 4.4.1) and JoMo (ext.) (see section 4.5.1) across different mechanisms of missingness namely MCAR and MAR. This chapter directly addresses the third objective specified in section 3.5 i.e. comparing performance of each MI method i.e. MICE (ext.) and JoMo (ext.) under different types of missingness in a three level model. The two approaches are compared against one another, and their performance is compared to results from performing Complete Case Analysis. The simulation study set-up is the same as explained in the previous chapters (see Section 4.2).

5.3 Methods

Our missing data methodology evaluated by simulations is subjected to the following four steps (Schouten, Lugtig, & Vink, 2018):

1. A complete data is simulated and considered as the population of interest.
2. Missing values are then introduced into the dataset to make it incomplete
3. The incomplete data are estimated using the competing missing data analysis strategies
4. Statistical inferences are obtained for the original (complete dataset), and after handling for missing values. A comparison of these inferences is used to measure the performance of each method (Salim, Mackinnon, Christensen, & Griffiths, 2008).

A similar approach is followed in our study. Multiple imputation procedures have an inherent random component within them, and use a pseudo-random number generator. In order to ensure reproducibility of results, we set the seed at the beginning of

the study.
Each simulated dataset is analysed in three ways, using the following:

1. Complete Case Analysis
2. Multiple Imputation Using Chained Equations i.e. MICE (ext.)
3. Joint Modelling i.e. JoMo (ext.)

In section 2.1.3, we defined the various types of missingness, namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). This thesis does not address the third mechanism of missingness (MNAR). Within each method, we considered three main situations where covariates at each level were set to 20% and 50% missingness (MCAR and MAR). This was done as the NFHS data saw on average atleast 20% missingness across variables in the dataset. We chose to introduce 50% missingness to see how well the method performs under extreme duress. 20% and 50% missingness were introduced in the variables as seen below:

1. 20% missing in X_{ijk}
2. 20% missing in both X_{ijk} and Z_{jk}
3. 50% missing in X_{ijk} , Z_{jk} and 20% missing in W_k

Missing values were introduced into the dataset under the MAR and MCAR mechanism, where the probability of missingness under MAR was determined by a logistic regression model (Lee & Carlin, 2017). Data generation for X_{2ijk} and Z_{2jk} are explained in Table 4.2.2. Probability of MAR in each of the variables has been explained in the previous chapter in section 4.2.1.

For simulating the MCAR mechanism, we created a binomial variable with probability of observing , equal to the required probability of missingness i.e. (20% or 50%), and set the corresponding values of X_{ijk}, Z_{jk} and W_k to missing. To briefly revisit our definition, we say that data is MCAR if the probability of missingness is the same for all units in the population, or $P(R|X) = P(R)$, where R takes values 0

Table 5.3.1: Definition and estimates of measures of performance used in the study

Performance Measure	*Definition*	*Estimators*
Absolute Relative Bias (ARB)	$\frac{E\lvert[\hat{\theta}]-\theta\rvert}{\theta}$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\frac{\lvert\hat{\theta}_i-\theta\rvert}{\theta}$
MSE	$E[(\hat{\theta}-\theta)^2]$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i-\theta)^2$
Probability coverage	$P(\hat{\theta}_{low}\leq\theta\leq\hat{\theta}_{upp})$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}1(\hat{\theta}_{low,i}\leq\theta\leq\hat{\theta}_{upp,i})$

Note: n_{sim} =1000

and 1 depending on if the corresponding variable is missing or observed respectively. For the purpose of this simulation, we created a binomial variable with probability of observing 1, equal to the required probability of missingness i.e. (20% or 50%), and set the corresponding values of X_{ijk}, Z_{jk} and W_k to missing.

To recap, the first simulated data was analysed before introducing missing values as a benchmark for expected results from any analysis procedure (see Table 4.2.3), and analysed after using cases with complete information. After imputation, the model was re-estimated for each imputed data set using the REML estimator of the *lmer*() function of the `lme4` R-package (Bates, Mächler, Bolker, & Walker, 2014) and the estimates were combined using Rubin's rules (Rubin, 1976).

5.3.1 Performance Measures

The performance of each of these imputation methods were compared using relative bias, mean squared error and probability coverage of the 95% CI (see Table 5.3.1). These are standard measures of performance were used as outcomes to compare the different methods against each other in estimating the true value of the parameter. When the target is an estimand, the most obvious performance measure to consider

Table 5.3.2: Performance Measures for the full data

Variable	*ARB*	*MSE*	*Coverage (%)*
Intercept	0.1814 (0.118)	0.3463 (0.435)	100%
$X_{ijk} - \bar{X}_{jk}$	0.1335 (0.113)	0.1501 (0.122)	100%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1338 (0.112)	0.1499 (0.122)	100%
$Z_{jk} - \bar{Z}_{k}$	0.1492 (0.126)	0.3006 (0.369)	100%
W_k	0.2842 (0.192)	0.5077 (0.432)	100%

is bias: the amount by which $\hat{\theta}$ exceeds θ on average (this can be positive or negative). Since we were more interested in the the magnitude of bias rather than direction, we use ARB (defined in 5.3.1) to capture the magnitude of bias in each method. Precision and coverage of confidence intervals will also be of interest (Morris, White, & Crowther, 2019). Coverage of confidence intervals is defined as the probability that a confidence interval contains the parameter. The distributions of imputed values and observed values are also compared to check if they overlap. We plot the parameters estimates against the iteration number to check for convergence. If the streams of parameter estimates are freely intermingled with one another and indicate no definite trends, then we can confirm the algorithm has converged. In addition, we also examine the distributions of the estimates and their SEs. Estimated values of parameters obtained from the analysis are compared to original values of parameters defined during data generation (see Section 4.2.1).

5.4 Results

Table 5.4.3 reports the performance of each method under scenarios with MCAR across the three levels, while Table 5.4.4 reports the performance of each method under scenarios with MAR across the three levels. We observe that when we have MAR in level 1 variable, here X_{ijk}, we see that MICE and JoMo both perform much better than Complete Case Analysis. MICE consistently performs better than JoMo across all parameters giving a coverage of more than 95% across all variables.

Table 5.4.3: Performance of imputation of missing data in a three level model (MCAR)

Variables	**Complete Case Analysis**			**MICE**			**JoMo**		
	ARB	*MSE*	*Coverage (%)*	*ARB*	*MSE*	*Coverage (%)*	*ARB*	*MSE*	*Coverage (%)*
Scenario 1: 20% MCAR in level 1 variables X_{ijk}									
Intercept	0.2963	0.8420	94.7%	0.2059	0.4208	94.5%	0.2033	0.4139	94.5%
$X_{ijk} - \bar{X}_{jk}$	0.9999	5.9000	0%	0.1335	0.1496	94.5%	0.1336	0.1502	89.1%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1337	0.14994	93.7%	0.1338	0.1496	96.8%	0.1337	0.1495	91.8%
$Z_{jk} - \bar{Z}_{k}$	0.1500	0.3013	95.4%	0.1505	0.3001	96.5%	0.1505	0.3005	95%
W_{k}	0.2890	0.5212	95%	0.2844	0.5073	95%	0.2850	0.5099	95.4%
Scenario 2: 20% MCAR in level 1 & 2 variables i.e. X_{ijk} & Z_{jk}									
Intercept	0.3013	0.8639	95%	0.2123	0.4352	100%	0.2027	0.4229	94.8%
$X_{ijk} - \bar{X}_{jk}$	1.0000	5.9002	0%	0.1417	0.1873	0%	0.1338	0.1503	0%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1339	0.1410	94.9%	0.1598	0.3252	0%	0.1337	0.1493	0%
$Z_{jk} - \bar{Z}_{k}$	0.1504	0.3017	95.1%	0.4133	1.2033	11%	0.1466	0.2789	87.5%
W_{k}	0.2889	0.5208	94.9%	0.2854	0.5107	100%	0.2837	0.5044	0%
Scenario 3: 50% MCAR in X_{ijk} & Z_{jk} and 20% MCAR in level 3 variable W_{k}									
Intercept	0.5101	2.3288	95.6%	0.2059	0.4219	100%	1.0173	7.6615	26.8%
$X_{ijk} - \bar{X}_{jk}$	0.9999	5.8998	0%	0.3567	0.9357	0%	0.1758	0.20222	69.8%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1340	0.1499	94.5%	0.3145	0.7728	0%	0.1392	0.1619	6%
$Z_{jk} - \bar{Z}_{k}$	0.1538	0.3056	94.6%	0.6998	3.0330	0%	0.1953	0.4033	91.8%
W_{k}	0.2981	0.5959	95.2%	0.2834	0.5023	0%	0.1755	0.2691	22.5%

Note: *ARB = Absolute Relative Bias (see Table 5.3.1 for definition)*

Table 5.4.4: Performance of imputation of missing data at multiple levels in a three level model (MAR)

Variables	**Complete Case Analysis**			**MICE**			**JoMo**		
	ARB	*MSE*	*Coverage (%)*	*ARB*	*MSE*	*Coverage (%)*	*ARB*	*MSE*	*Coverage (%)*
Scenario 1: 20% MAR in level 1 variables X_{ijk}									
Intercept	3.4344	64.6510	0%	0.2078	0.4309	95.4%	0.2078	0.4308	95%
$X_{ijk} - \bar{X}_{jk}$	0.9993	5.8932	0%	0.1334	0.1496	98.3%	0.1337	0.1502	97.5%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1335	0.1495	91.7%	0.1338	0.1496	97.8 %	0.1336	0.1488	83.8%
$Z_{jk} - \bar{Z}_{k}$	0.1495	0.2998	94.7 %	0.1500	0.2997	97.3%	0.1498	0.2988	96.4%
W_k	0.1504	0.1772	9%	0.2846	0.5082	95%	0.2846	0.5081	94.4%
Scenario 2: 20% MAR in level 1 & 2 variables i.e. X_{ijk} & Z_{jk}									
Intercept	3.0703	51.6601	0%	0.2098	0.4333	100%	0.2158	0.4497	95.1%
$X_{ijk} - \bar{X}_{jk}$	0.9983	5.882	0 %	0.1429	0.2082	0%	0.1340	0.1505	100%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1334	0.1489	84.5%	0.2037	0.4259	0%	0.1349	0.1422	0.01%
$Z_{jk} - \bar{Z}_{k}$	0.1491	0.2885	91.9%	0.1487	0.2416	99.5%	0.1421	0.1825	4%
W_k	0.1565	0.1791	17.9 %	0.2846	0.5072	100%	0.2849	0.5068	94.3%
Scenario 3: 50% MAR in X_{ijk} & Z_{jk} and 20% MAR in level 3 variable W_k									
Intercept	4.9025	135.3929	0%	0.0292	0.4145	100%	2.9114	47.2239	2.6%
$X_{ijk} - \bar{X}_{jk}$	0.9987	5.8858	0%	0.1901	0.7104	0%	0.1690	0.1895	100%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1336	0.14857	86.5%	0.1165	0.4335	0%	0.1352	0.1400	98.2%
$Z_{jk} - \bar{Z}_{k}$	0.1503	0.2734	90%	4.1730	61.4779	0%	0.1398	0.1496	51%
W_k	0.1592	0.2283	32.9%	0.2784	0.5009	100%	0.4975	1.6974	0.03%

Note: *ARB = Absolute Relative Bias (see Table 5.3.1 for definition)*

We observe a change in this performance as we include a combination of MAR in level 1 and level 2 variables (here, X_{ijk} and Z_{jk}). We see that MICE and JoMo show lower values of MSE and ARB, compared to Complete Case Analysis, however coverage of both imputation methods are quite low when estimating for derived variables. We see that probability coverage of the 95% CI for the parameters for variables $X_{ijk} - \bar{X}_{jk}$ and $\bar{X}_{jk} - \bar{X}_k$ is 0% while imputing using MICE and the coverage of variables $\bar{X}_{jk} - \bar{X}_k$ and $Z_{jk} - \bar{Z}_k$ is $< 5\%$ when imputing using JoMo. We produce coverage plots to investigate these anomalies further (see Figure 5.4.1).

The red and blue dots of Figure 5.4.1) are respectively the lower and upper limits of the CI, and the black line shows the true value of the parameter. Across all three scenarios, with varying proportions of MCAR introduced at each level, we observe that multiple imputation methods perform very differently as compared to the results obtained under MAR (see Table 5.4.4).

We observe lower values of MSE and ARB across all three scenarios when Complete Case Analysis was used. Probability coverage of 95% CIs is also higher that what was seen in Table 5.4.4. However, we observe a few variations when multiple imputation is used to handle for MCAR. We observe low values of bias and MSE, and high probability coverage of CIs for both MICE and JoMo when MCAR is observed at level 1 alone (X_{ijk}).

These results deteriorate when missing values are seen in combination across all three levels in the dataset (scenarios 2 and 3). Table 5.4.4 shows low values of bias and MSE, but also lower values of coverage of CIs. This reason for these results are explored further in section 5.6). As we move down the rows in Table 5.4.3, we see a steady improvement in performance of Complete Case Analysis compared to MI methods. Complete Case Analysis has been established as an unbiased method of analysis of missing data when data is MCAR. However, it is interesting to see its steady performance when missing values are present across all levels in the data. Values of MSE and absolute relative bias are closer to zero for even up to the third

scenario. However, we see a 0% coverage for the variable $X_{ijk} - \bar{X}_{jk}$ across all three scenarios. My intuition is that this lack of coverage can be attributed to the derived nature of the variable. In this scenario missing values were imputed in the variable X_{ijk}, and then the variable $\bar{X}_{jk}$ was recalculated using passive imputation, using the newly imputed values of X_{ijk}. This process has might have created a shift in the distribution, resulting in 0% probability coverage of the confidence interval. This interpretation is further strengthened by the low ARB and MSE values observed for these variables.

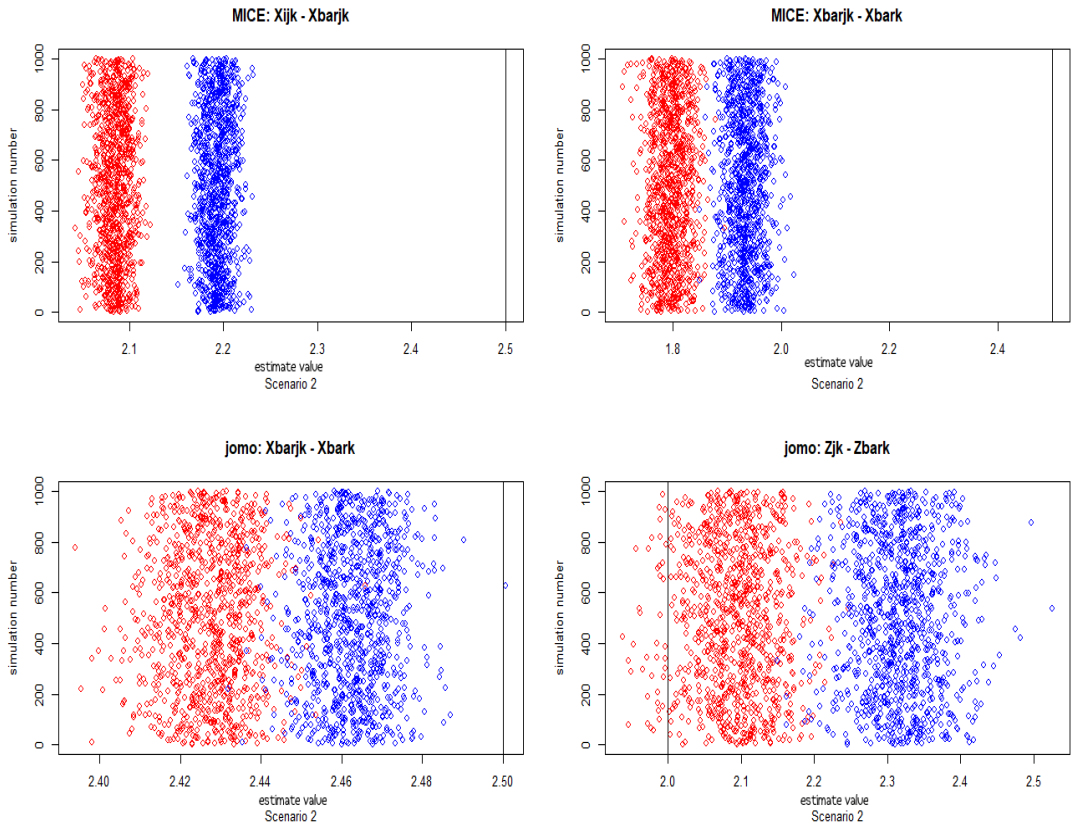


Figure 5.4.1: Highlighting coverage issues (from Table 5.4.4) in MICE and JoMo for MAR observed in Level 1 and Level 2 variables (combined)

We observe that while the estimates are precise (as shown by the narrow width of

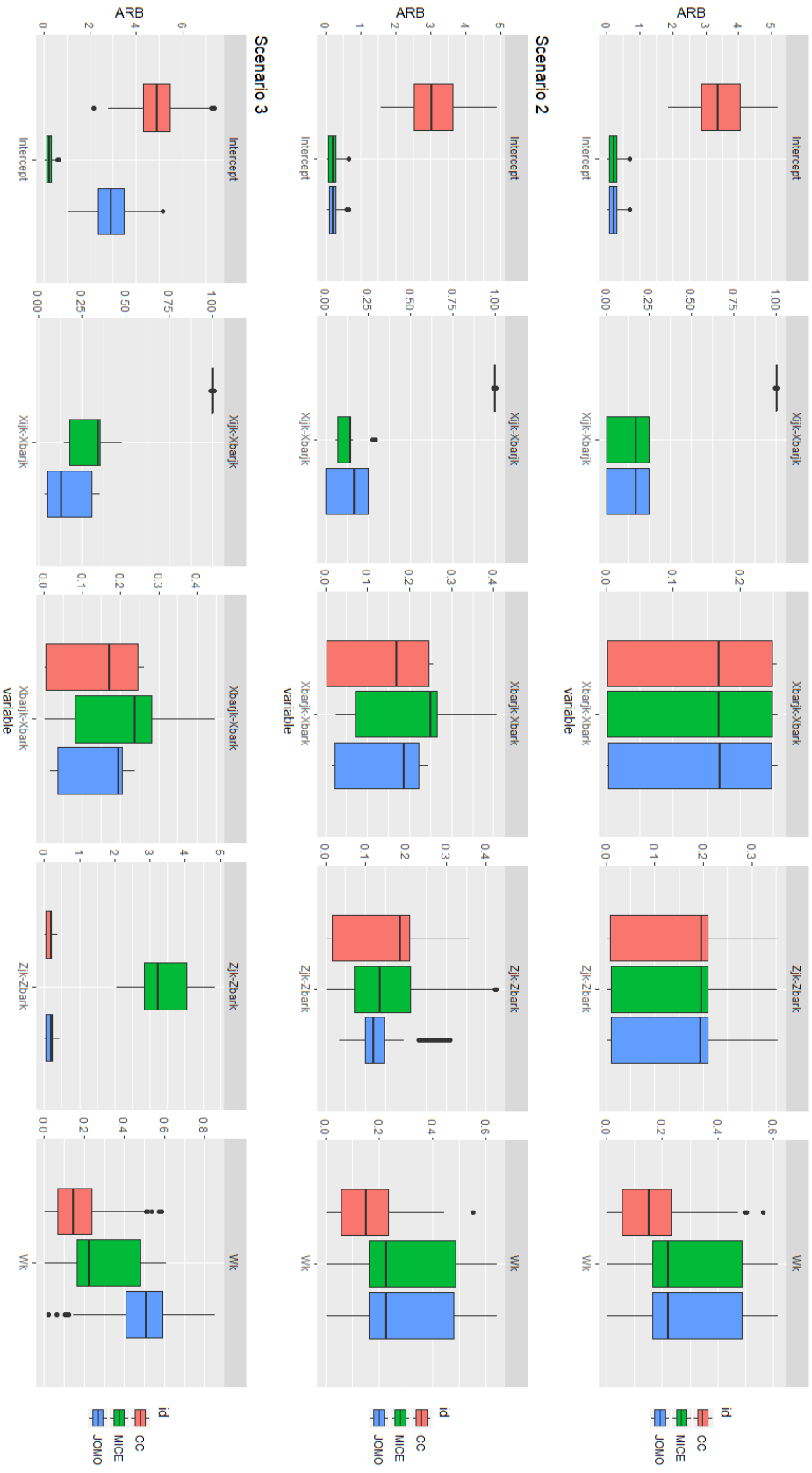


Figure 5.4.2: Distribution of Absolute Relative Bias obtained from 1000 simulations across 3 methods of missing data analysis

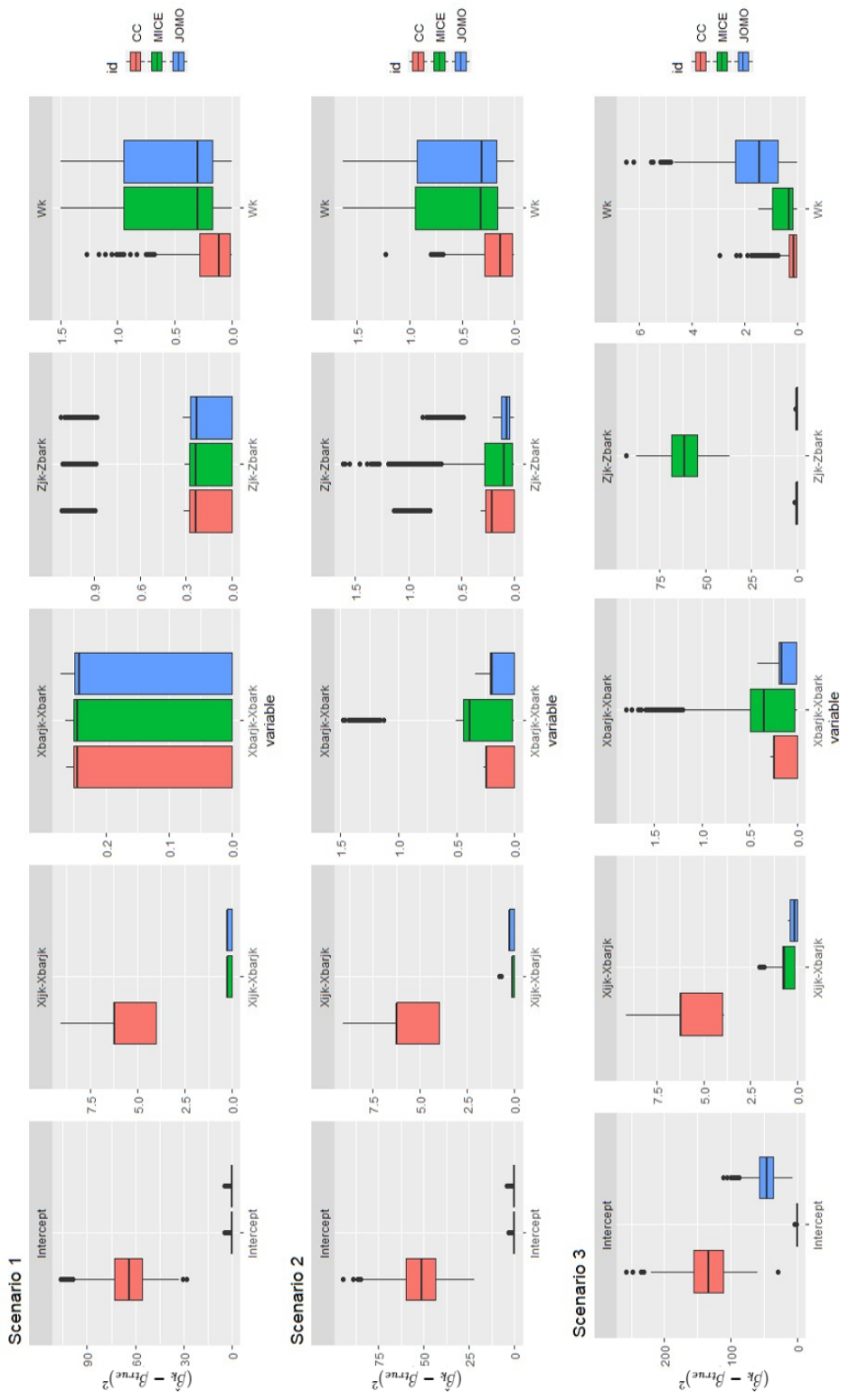
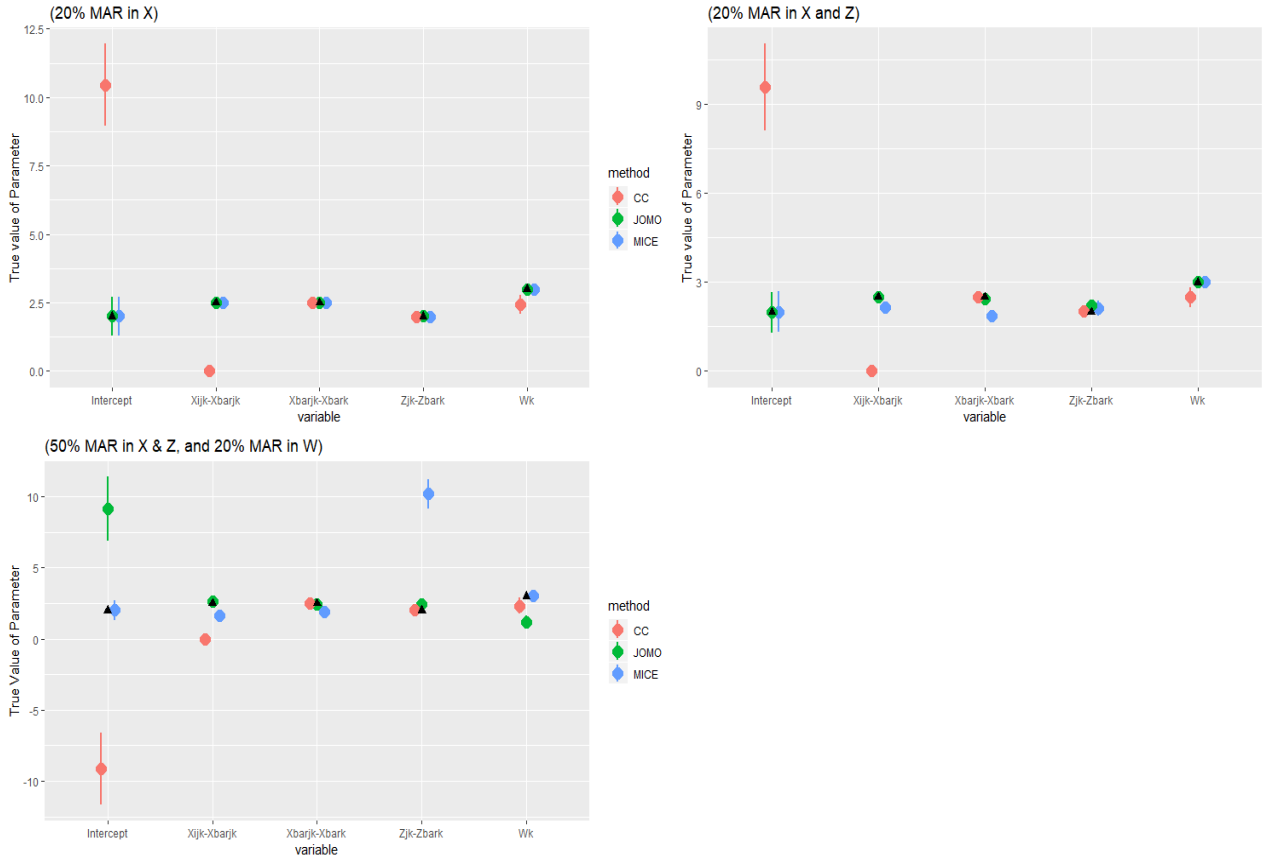


Figure 5.4.3: Distribution of $(\hat{\beta_k} - \beta_{true})^2$; (k indexing simulations)obtained from 1000 simulations across 3 methods of missing data analysis

the intervals), the true value of the parameters are not captured in these confidence intervals (shown in Figure 5.4.1), as shown by the black line. It is curious to note that after passive imputations, the 95% CIs fails to capture the true values of the parameter for the derived variables, i.e. $X_{ijk} - \bar{X}_{jk}$, $\bar{X}_{jk} - \bar{X}_k$ and $Z_{jk} - \bar{Z}_k$. It is crucial to note that these derived variables were recalculated using passive imputation. This means that the derived variables, were recalculated using the imputed values of both X_{ijk} and Z_{jk}. These results continue to echo for imputations done in scenarios with higher proportions of missingness in the dataset (i.e. Scenario 2 and 3). One can speculate that the "non-coverage" of the CIs could be due to the *slight but visible* shift of the distribution of imputed values from the truth, as observed by the narrow difference between the true parameter value and the upper limit of the CIs (see Figure 5.4.1). Under coverage can be expected due to the following reasons, (a) Bias $\neq 0$, (b) Model SE $<$ Empirical SE, (c) The distribution of $\hat{\theta}$ is not normal and the intervals are constructed assuming normality, or (d) $\hat{Var}(\theta_i)$ is too large(Morris et al., 2019). Morris states that under coverage due to bias will reduce as n_{obs} increases. However, this is not tested till Chapter 6. A comparison of the model based and empirical SE was tested preliminarily in this thesis, but there wasn't sufficient evidence to suggest significant difference between the two standard errors, and was not included in this thesis.

In order to investigate this further, we calculate 95% credible intervals for each scenario. The results of this are shown in Figure 5.4.4. The black triangles show the true parameter values.

Figure 5.4.4: Credible Intervals obtained from 1000 simulations across 3 methods of missing data analysis



In addition, we also compare the distribution of the absolute relative bias and MSE for each scenario. Figure 5.4.2 shows a comparison of the distribution of the estimates of absolute relative bias obtained across the three missing data methods. Comparing the distribution of relative bias obtained for MICE for Scenarios 2 and 3, we see that imputation using MICE produces low values of relative bias (approximately zero) across the three scenarios. A deviation from this is observed in scenario 3 where MICE shows greater bias than the other two methods. JoMo shows strong results across all scenarios, with lower values of bias in estimation of parameters. We see an anomaly here where bias in the intercept values for JoMo is greater than MICE (see Figure 5.4.2), yet the method fares better than Complete Case Analysis across the board. Across most facets in the scenarios, we observe that the distribution of the

bias is quite skewed, with a bulk of the bias values concentrated between 0 and 0.15, indicating low values of bias in estimation of the relevant parameters.

Discussion

When analysing multi level datasets with missing values present across all levels in the dataset, multilevel multiple imputation is a widely applicable and reasonably accurate technique to handle for missing data. One faces the challenge of deciding the method of multiple imputation that needs to be used in order to perform imputation. In this chapter, we have compared the results obtained by performing MI using both MICE and JoMo. Regardless of the choice of imputation method used, there are certain principles to bear in mind.

- Account for data structure in the analysis model & the imputation model.
- The Imputation model should be compatible with the analysis model.
- Derived variables need to be recalculated using the imputed values of related variables.

In this chapter, we evaluated the simulation performance of MI in three level models when there are missing values across all levels in the dataset. The simulation employed the Gelman and Hill approach of analysis in developing the imputation model.

The construction of the imputation model closely follows the recipe defined in Table 4.4.4. As per the algorithm, we identify the target variables with missing values, i.e. X_{ijk}, Z_{jk} or W_k depending on the corresponding scenario. The choice of variables used in the imputation model closely follows the algorithm, by including all cluster means of lower level variables, and excluding any variable that does not occur in the analysis model. Finally, any terms involving the target variable, was not included in the imputation model, but was passively imputed in the algorithm. This approach was followed in both imputation procedures.

The imputation model for higher level units were constructed by using aggregate measures of lower level variables. In addition, derived variables (or linear transformations of variables) that are commonly used in models were calculated using Passive Imputation.

We considered various combinations of MAR in each level in the dataset and gradually increased the degree of missingness observed across the levels. We observed that across most scenarios, MI methods performed better than Complete Case Analysis. JoMo and MICE both yield good regression estimates, even though there were some inconsistencies observed across the variables in the dataset, especially with respect to coverage of confidence intervals. Our analysis demonstrated that despite these inconsistencies, both multiple imputation methods provide estimates with minimum bias and mean square error estimates, when imputing for missing values in higher level units.

In terms of user control, we observe that MICE allows us to define an imputation model for each variable with missing data. This gives the researcher more freedom in terms of including variables into the model. This observation is consistent with the definition of how MICE functions (explained in section 4.4). This is not possible in JoMo as it defines a multivariate joint distribution (as explained in section 4.5) for all variables in the model. In our simulations, this did not present as a limitation and JoMo produced estimates with low bias and mean square error. However, we do worry that this joint distribution may not show consistent performance across all research examples, and using JoMo to impute in such cases, may force a joint distribution that does not exist. Thus the decision of choosing the right imputation method is not one to be taken lightly.

5.5 Comparing performance of Multilevel MI (MCAR versus MAR)

We now refocus our attention to the performance of MICE and JoMo under all three scenarios of MAR and MCAR. On comparing the distribution of bias (see Figure 5.5.5) and MSE (see Figure 5.5.6), we see that as the proportion of missingness increases in the data, MICE shows lower levels of bias compared to JoMo. While the difference is not very obvious in scenario 1, the shift becomes more noticeable as we move across scenarios. With respect to MSE values, we see both methods have reasonably low levels of MSE values except in the third scenario (which is an extreme case). This is promising as we deduce that MICE and JoMo consistently show better results than Complete Case Analysis when missing data is present across all levels in the data, regardless of the mechanisms of missingness. This observation is key, as the scenarios demonstrated here compare moderate, severe and extreme levels of missingness that were introduced across various levels in the dataset.

As in Section 5.4, we observe that coverage for both MICE and JoMo are low compared to Complete Case Analysis, when values is MCAR. As in Figure 5.4.4, we make use of credible intervals to give us a more holistic understanding of the range of the confidence interval obtained. Given that coverage of 95% CIs is quite high for Complete Case Analysis, we make compare credible intervals only for the two imputation methods used (see Figure 5.5.7).

5.5.1 Results

Here, we compare the performance of MICE (ext.) and JoMo (ext.) when data is MCAR and MAR. As mentioned at the start of the chapter, this is directly aimed at addressing the third objective which discusses the performance of these extended methods of imputations to handle various mechanisms of missingness in three level datasets.

On further investigation, the low coverage values of MICE and JoMo were probably

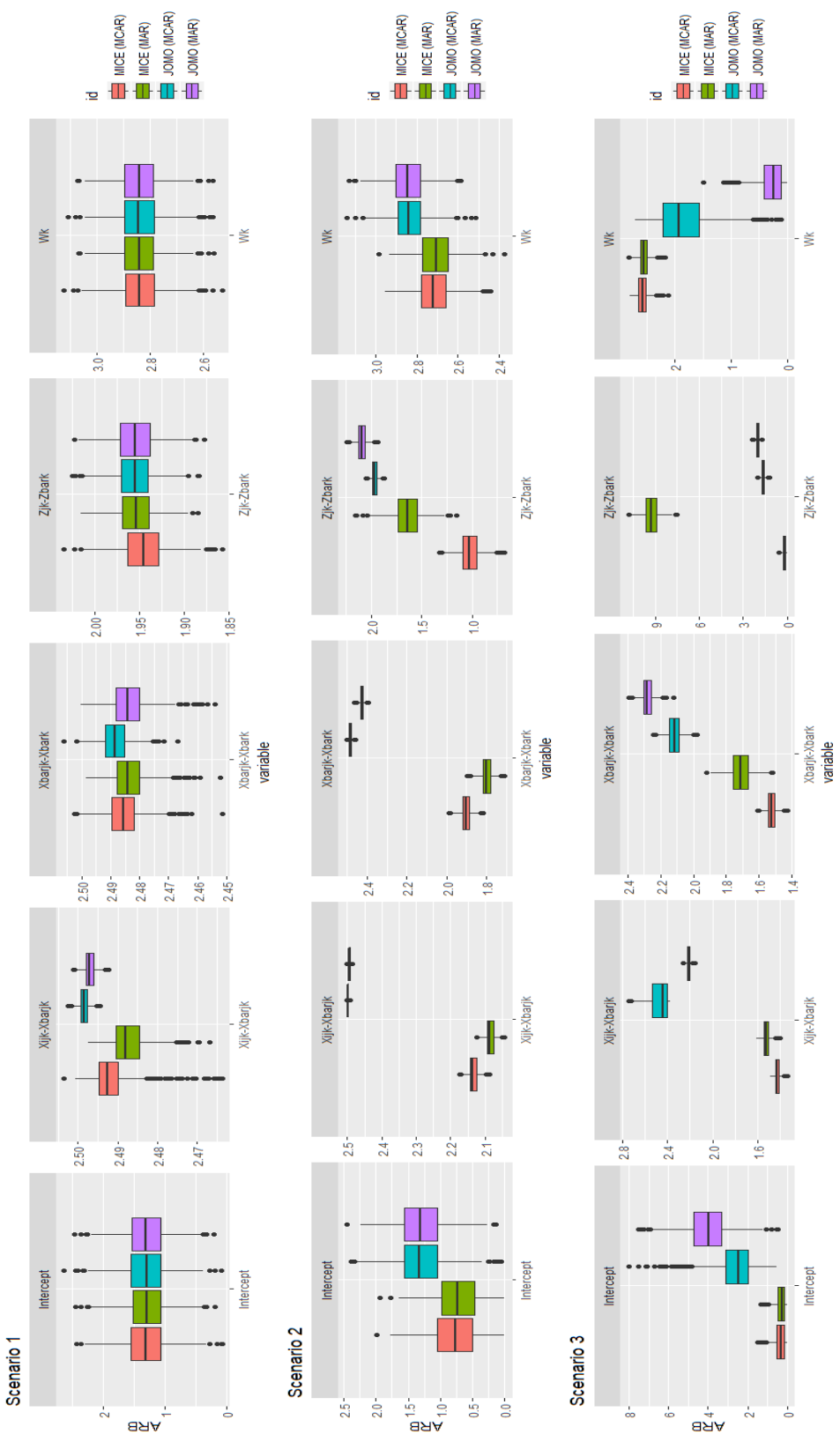


Figure 5.5.5: Distribution of Absolute Relative Bias obtained from 1000 simulations across different mechanisms of missingness for MICE and JoMo

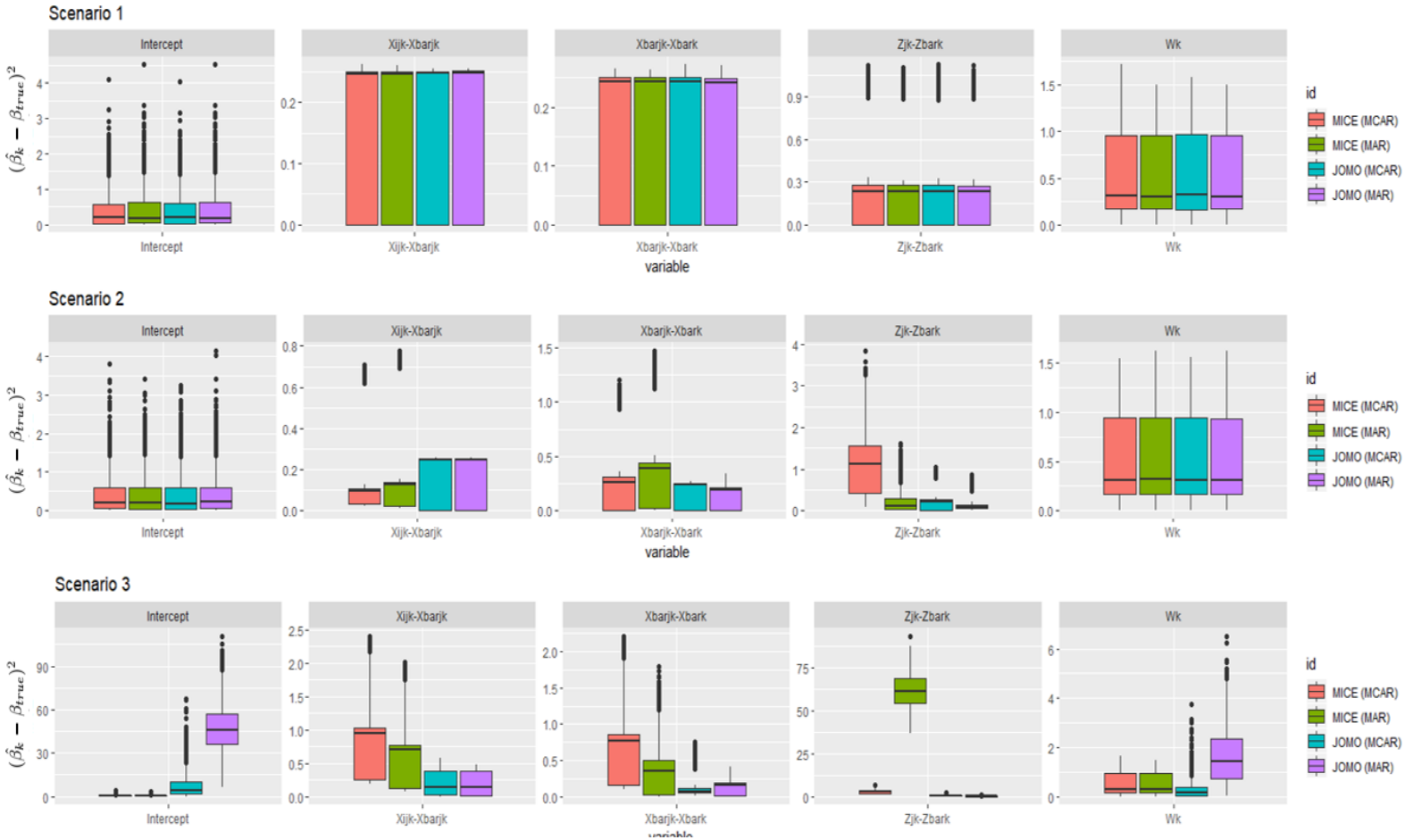


Figure 5.5.6: Distribution of $(\hat{\beta}_k - \beta_{true})^2$; (k indexing simulations) obtained from 1000 simulations across different mechanisms of missingness for MICE and JoMo

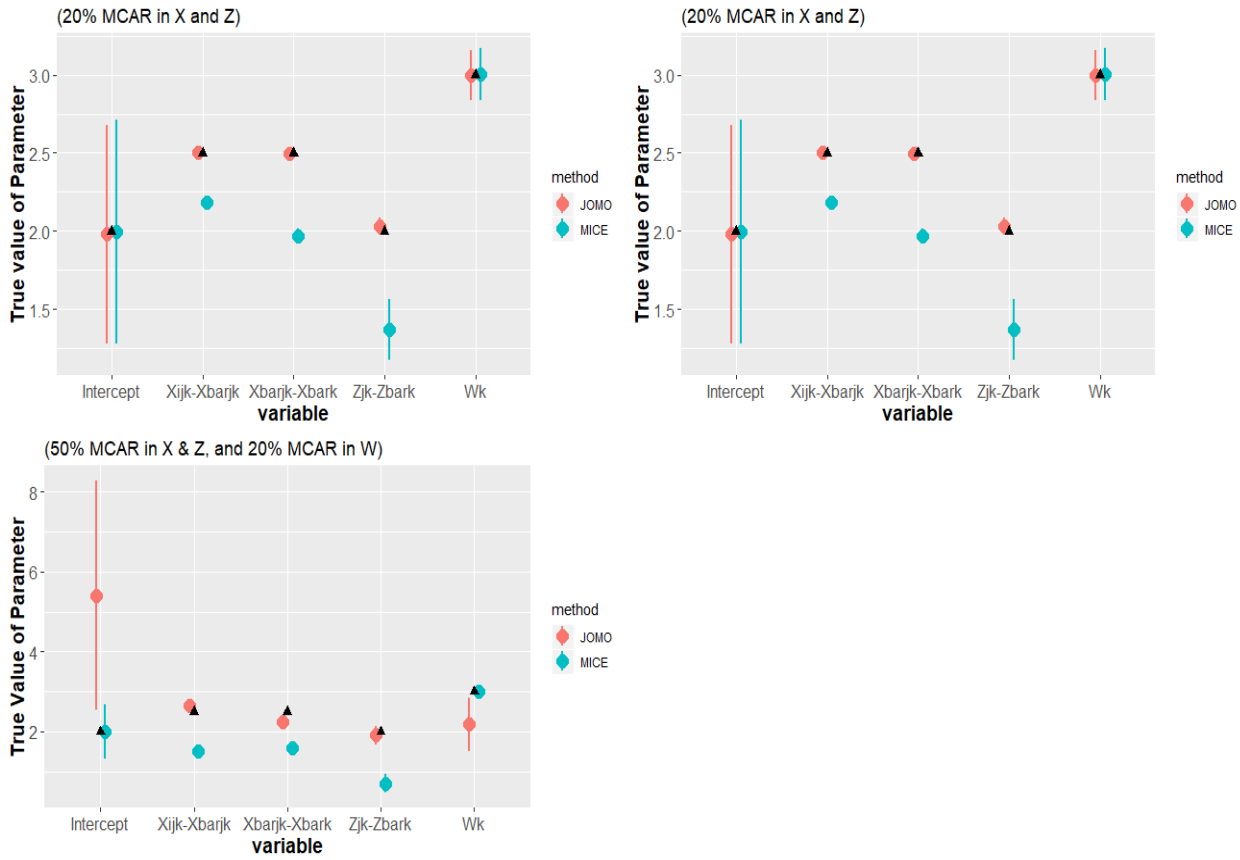


Figure 5.5.7: Credible Intervals obtained from 1000 simulations across 3 scenarios for both MICE and JoMo

caused by a small shift in the distribution of the estimates obtained post imputation. This is further ascertained from Figure 5.5.7; where we observe that the credible intervals for both MICE and JoMo are quite small and clustered around the true value (denoted by the black triangle).

Overall, the results are promising when we compare the performance of the imputation methods under different mechanisms of missingness after taking account of the data structure in the imputation models.The next section zooms in on the results obtained for coefficients of derived variables to investigate some inconsistencies that were highlighted during the simulations. It aims to verify the choice of methods used, further substantiate the performance of the methods and answer some key

questions proposed through the simulations.

5.6 Zooming in on the gaps identified from the Simulation Study

5.6.1 Introduction

While the results from the analysis (see section 5.4) helped better understand the performance of the methods of imputations and provided relevant and quality information on the each method of imputation, it also helped identify few gaps in our understanding of their performance. We see in Figures 5.4.2 and 5.4.3 that the three methods of analysis for missing data performed equally and reported similar MSE and Relative bias estimates. This raises two questions; (1) Are the three methods of analysis comparable for handling missing data occurring only in level 1 variables? (2) Are the three methods of analysis comparable because there is only 20% missingness across the dataset in the level 1 variable? The latter would indicate that the impact of missingness in the dataset is insignificantly small, and so Complete Case Analysis shows as much bias as MICE and JoMo.

Before we begin this investigation, we recall that in this thesis, we identify variables with the subscripts i,j and k to signal to the level at which the variable is captured. For example, X_{ijk} is a variable captured at the lowest level (Level 1) where (i) refers to the number of units at the lowest level, (j) refers to the second level (Level 2) and (k) refers to the third level (Level 3). Here, the subscript (i) denotes the individuals, (j) denotes the households and (k) denotes the districts. Readers should note that throughout this thesis, the covariates measured at the lowest level are signalled to by the letter X, the covariates at the second level are denoted by the letter Z and the covariates at the highest levels are denoted by the letter W. Outcome variables are denoted by the letter Y.

The second aspect that was highlighted by the results of the study were the non-coverage of CIs for derived variables by both MICE and JoMo. As discussed above, derived variables $X_{ijk} - \bar{X}_{jk}$, $\bar{X}_{jk} - \bar{X}_k$ and $Z_{jk} - \bar{Z}_k$ were recalculated by using the im-

puted values of X_{ijk} and Z_{jk}. This method of handling for derived variables is known as **Passive Imputation**. It is also commonly referred to as the Impute → Transform method (mentioned in Section 2.3.9) which is indicative of the its process. The (non) coverage plots in Figure 5.4.1 highlighted some key dependencies that are carried over in the process resulting in the *slight shift* of the confidence interval from the true parameter. While passive imputation is considered a useful method to handle derived variables in the model, the results from the simulation raise the questions (3) Are there other methods to handle derived variables? (4) Is non-coverage of confidence intervals a consequence of having derived variables in the model?

To further investigate and answer these four questions, we propose the following possibilities.

1. Increase proportion of missingness from 20% to 50% (extreme), and compare the performance of MICE, JoMo and Complete Case Analysis with the results from table 5.4.4.

2. Calculate probability coverage of imputation of CIs for derived variables using alternate methods (e.g. JAV)

5.6.2 Increase proportion of missingness from 20% to 50%

In scenario 1 of Table 5.4.4, we observed that the three methods of imputation performed equally when missing values were present at only level 1 variable X_{ijk}. We hypothesised that this could be explained by having only 20% missing values at the lowest level variable. In order to test this hypothesis, we now raised this proportion to 50% MAR in X_{ijk}. As in the previous scenario, probability of MAR in X_{ijk} was determined by model $\frac{e^{X_s+\beta Y_s}}{1+e^{X_s+\beta Y_s}}$; where $Y_s = \frac{(Y-E(Y))}{SD(Y)}$ & $X_s = \frac{X-E(X)}{SD(X)}$ We follow the same method of evaluating simulations as described in section 5.3. Since simulations are heavily based on random number generation, in order to ensure that the same dataset was being compared across scenarios, we specified the same seed in all scenarios. In order to make the results comparable, we compare the performance of the each of these

methods i.e. Complete Case Analysis, MICE (ext.) and JoMo (ext.); using absolute relative bias, MSE and probability coverage of CIs. Definition and calculation of the estimates of these measures are described in Table 5.3.1.

The results from Table 5.6.5 provides a two-dimensional insight into the performance of the three methods. First, they allows us to compare the performance eof each method against itself when the degree of missingness increases from 20% to 50%. Second, they allow us to compare the three methods against each other to compare their performance when the proportion of missing values increases at the lowest level variable. To address the former, we see that the absolute relative bias and MSE values in column one and two increase substantially for the intercept estimates; going up from 3.443 to 4.867, and 64.651 to 132.39 respectively. It is positive to note that even in extreme missing cases scenarios (50% MAR) at level 1 variables, MICE and JoMo continue to perform well reporting low values of MSE and relative bias across all variables. Comparing these results to Scenario 3 in Table 5.4.4, which also showed an extreme scenario of 50% missingness in X_{ijk} coupled with varying proportions of missingness in other variables, it is positive to note that the lack of coverage for $X_{ijk} - \bar{X}_{jk}$ and $\bar{X}_{jk} - \bar{X}_{k}$ i.e (0% for both), can possibly be attributed to the combination of missingness across the three levels rather than the 50% missingness observed at level one alone.

From Table 5.6.5, we see that probability coverage of imputation of CIs for MICE is well above 95% for all variables in the model; however we see this falls for coverage of variable $\bar{X}_{jk} - \bar{X}_{k}$ when JoMo is employed. This suggests a slight drop in performance of JoMo when proportion of missingness increases in level 1 variables. Comparing to Scenario 3 from Table 5.4.4 which also observed 50% MAR in X_{ijk} along with varying proportions of missingness across other variables in the dataset, we see improved performance of JoMo across all measures, indicating that the differences may be due to the combination of missingness across all variables rather than the extreme scenario observed at level 1 alone. Finally, we compare the distribution of MSE and Absolute relative bias for these two scenarios (20% and 50% MAR in X_{ijk}).

Table 5.6.5: Results of imputation after increasing MAR in X_{ijk} from 20% to 50%

Variables	**Complete Case Analysis**			**MICE**			**JoMo**		
	ARB	*MSE*	*Coverage*	*ARB*	*MSE*	*Coverage*	*ARB*	*MSE*	*Coverage*
Scenario 1.1: 20% MAR in level 1 variable X_{ijk}									
Intercept	3.4344	64.6510	0%	0.2078	0.4309	95.4%	0.2078	0.4308	95%
$X_{ijk} - \bar{X}_{jk}$	0.9993	5.8932	0%	0.1334	0.1496	98.3%	0.1337	0.1502	97.5%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1335	0.1495	91.7%	0.1338	0.1496	97.8%	0.1336	0.1488	83.8%
$Z_{jk} - \bar{Z}_{k}$	0.1495	0.2998	94.7%	0.1500	0.2997	97.3%	0.1498	0.2988	96.4%
W_k	0.1504	0.1772	9%	0.2846	0.5082	95%	0.2846	0.5081	94.4%
Scenario 1.2: 50% MAR in level 1 variable X_{ijk}									
Intercept	4.8675	132.39	0%	0.2078	0.4309	95.4%	0.20778	0.4309	94.7%
$X_{ijk} - \bar{X}_{jk}$	0.9991	5.8900	0%	0.1341	0.1503	98.2%	0.1371	0.1529	100%
$\bar{X}_{jk} - \bar{X}_{k}$	0.1337	0.1495	90.5%	0.1343	0.1503	98.7%	0.1343	0.1437	69%
$Z_{jk} - \bar{Z}_{k}$	0.1503	0.3005	94.8%	0.1502	0.2994	96.3%	0.1575	0.3131	94.2%
W_k	0.1647	0.2093	37.1%	0.2846	0.5082	95%	0.2846	0.5082	94.4%

5.6.3 Probability coverage of CI for derived variables using Just Another Variable (JAV) Approach

In section 5.6, we note that there might be some key dependencies between the imputed variables that are carried over in the process of passive imputation, resulting in the non-convergance of confidence intervals (see Figure 5.4.1). The method used previously was Passive Imputation, however we posed the question of using alternate methods to handle derived variables to improve coverage. In this section, we compare results obtained from using Passive Imputation approach versus an alternate approach, called Just Another Variable (JAV) approach, in the context of multilevel multiple imputation.

In Figure 5.4.1, we observed that low coverage probabilities were reported for MICE and JoMo when estimating the parameters for the derived variables $X_{ijk} - \bar{X}_{jk}$, $\bar{X}_{jk} - \bar{X}_k$ and $Z_{jk} - \bar{Z}_k$. On closer inspection of the coverage plots, we observed that while confidence intervals for the parameters were narrow, they were not accurately capturing the true parameter values. In addition, the difference between the upper (or lower limits) and the true value was very small ($\leq$ 0.5 units). This leads us to believe that a certain procedure in the imputation was resulting in a *slight but visible* shift in the distribution of the values. This led us to further inspect the process of handling derived variables. Across literature (with the sole exception of a study by von Hippel (2009), passive imputation has been defined as the standard process to handle derived variables. Passive imputation is also referred to as the Impute $\rightarrow$ Transform process. It involves imputing for the variable with missing values and then recalculating the derived variables using these imputed values. The choice of imputation method for the derived variables, is crucial as we observed 20% missingness across level 1 and level 2 variables. It is recognised that when the imputation model includes squares and interactions in the imputed data, the recommendation is to impute them in their raw form and then transform them. von Hippel (2009) and van Buuren (2018) have both demonstrated that passive imputation yields good regression estimates. However, their work primarily focused on non-linear trans-

formations such as squares and interactions, whereas our research focuses on linear transformation. Now, $\bar{X}$ is a linear transformation of X since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. An additional challenge is that the data is multilevel. Following general recommendations von Hippel (2009), we calculated $\bar{X}_{jk}$ and $\bar{X}_{k}$ before imputation and then recalculated it after imputation using the imputed values. This was done at between each cycle in the FCS process. This is consistent with van Buuren (2018) and Vink's advice on their blog "mice: Passive imputation and Post-processing". [1]

To investigate whether passive imputation is the best approach, we consider an alternate method of imputing derived variables known as "Just Another Variable (JAV)" approach. In the JAV approach, Y_{ijk}, X_{ijk} and $\bar{X}_{jk}$ are assumed to be jointly normally distributed, and each function of the variable with missing data is treated as an unrelated variable. Seaman et al.(2012) demonstrated that JAV performed better than passive imputation for linear regression with a quadratic or interaction term. It's performance in comparison to passive imputation for multilevel regression models was not tested. The study did not investigate the methods performance where additional covariates were involved.

In Section 5.6.1, we posed two questions i.e. *Are there other methods to handle derived variables?*, and *Is non-coverage of confidence intervals a consequence of having derived variables in the model?*. To answer these questions, we compare the coverage of confidence intervals for derived variables. This allows us to compare the plots from Figure 5.4.1 that were obtained from passive imputation with the coverage plots obtained from imputation of derived variables using the JAV approach. With the exception of $X_{ijk} - \bar{X}_{jk}$, for both MICE and JoMo, JAV approach for handling cluster means showed excellent coverage of confidence intervals for all other derived variables in the model. For the variable $X_{ijk} - \bar{X}_{jk}$, we also observe that the difference between the upper limit of the CIs and the true value seems to have increased (going from 0.5 units to 2.5 units) when comparing with Figure 5.4.1. In addition, the CI for the $X_{ijk} - \bar{X}_{jk}$ also seems to center around zero implying that the imputed values for X_{ijk}

[1] https://www.gerkovink.com/miceVignettes/Passive_Post_processing/Passive_imputation_post_processing.html

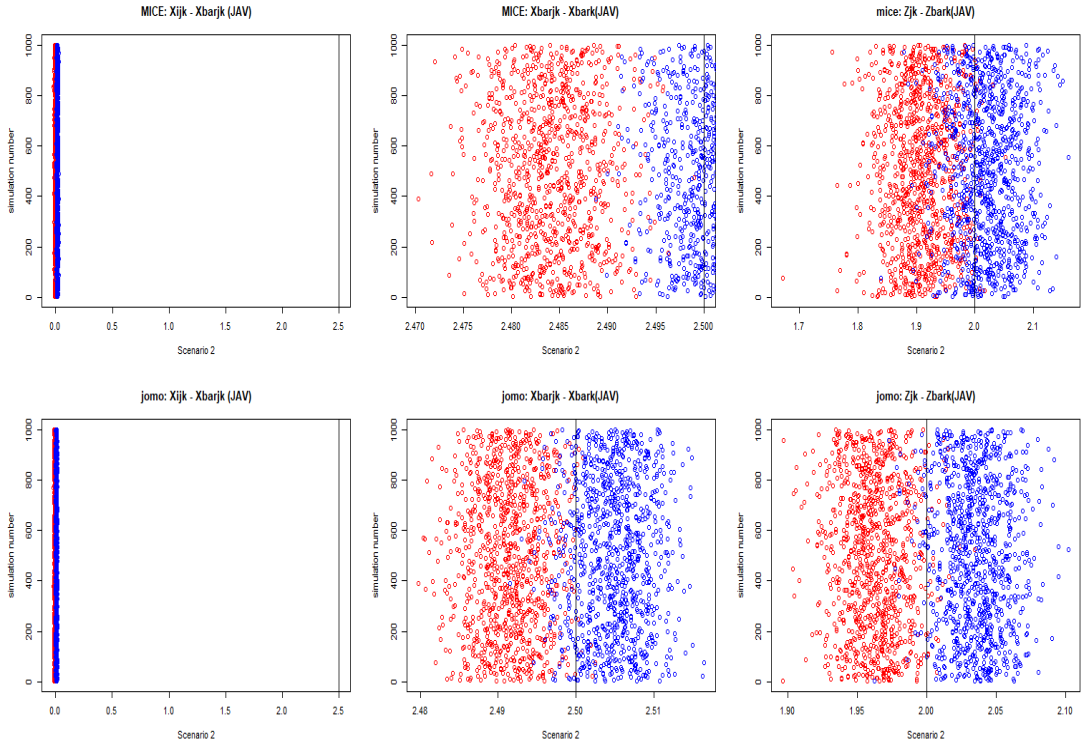


Figure 5.6.8: Highlighting coverage of CI for parameter estimates of derived variables using JAV in combination with MICE and JoMo for MAR observed in Level 1 and Level 2 variables (combined)

and $\bar{X}_{jk}$ would be the same. This means that the cluster means are very close to each individual value. This would imply that within cluster variation at level 1 was approximately equal to 0. This effect is not visible for level 2 derived variables $Z_{jk} - \bar{Z}_k$ which shows around 99.5% coverage of CIs. However, performance cannot be measured by coverage alone. To this end, we compare performance of both methods of imputation for derived variables using our standard measures of performance. The results of this are presented in Table 5.6.6.

Table 5.6.6: Comparing performance of parameter estimates imputing derived variables using Passive Imputation versus JAV approach

20% MAR in level 1 & 2 variables i.e. X_{ijk} & Z_{jk}						
Variables	**MICE**			**JoMo**		
	ARB	*MSE*	*Coverage*	*ARB*	*MSE*	*Coverage*
Scenario 2.1: **(Passive Imputation)**						
Intercept	0.2098	0.4333	100%	0.2158	0.4497	95.1%
$X_{ijk} - \bar{X}_{jk}$	0.1429	0.2082	0%	0.1340	0.1505	100%
$\bar{X}_{jk} - \bar{X}_k$	0.2037	0.4259	0%	0.1349	0.1422	0.01%
$Z_{jk} - \bar{Z}_k$	0.1487	0.2416	99.5%	0.1421	0.1825	4%
W_k	0.2846	0.5072	100%	0.2849	0.5068	94.3%
Scenario 2.2: **(Just Another Variable)**						
Intercept	3.0807	51.9765	0%	3.009	49.6642	0%
$X_{ijk} - \bar{X}_{jk}$	0.9971	5.8678	0%	0.9996	5.8955	0%
$\bar{X}_{jk} - \bar{X}_k$	0.1336	0.1487	62.9%	0.1336	0.1495	87.5%
$Z_{jk} - \bar{Z}_k$	0.1643	0.3319	66%	0.1502	0.3014	93.6%
W_k	0.1533	0.1752	16%	0.1589	0.1852	21.3%

Note: *ARB = Absolute Relative Bias (see Table 5.3.1 for definition)*

Table 5.6.6 provides a good example on why one must not rely solely on graphs to interpret performance of methods. On comparing the estimates obtained from both methods i.e. Passive Imputation and JAV approach, we see that when values are imputed using the latter, the results are more biased. The coverage slightly improves when JAV is employed, although this is not consistent across all variables in the regression model. As in Figure 5.4.4, we plot the 95% credible intervals for both MICE and JoMo. In Figure 5.6.8, unlike the previous analysis (see Tables 5.4.3 and 5.4.4), Complete Case Analysis is not represented here as the focus of this section is on the imputation techniques for handling derived variables. In addition, the comparison is not applicable when one chooses to use Complete Case Analysis, as the process of recalculating derived variables post imputation does not exist.

5.7 Summary

In this chapter, we highlighted some key questions that were raised through our simulation study results. We concluded that imputation methods such as MICE and JoMo continue to perform extremely well, producing results with low bias and MSE values, and excellent coverage of confidence intervals (>95%) even in extreme scenarios such as 50% MAR at the lowest level variables. We also concluded that Passive imputation continues to remain a better approach when handling derived variables such as cluster means and deviations from cluster means from imputations, even in three level settings. This is concurrent with the existing literature on passive imputation for independent observations (van Buuren, 2018; von Hippel, 2009).

The next chapter looks at the impact of varying cluster sizes in three level data structures on multiple imputation. Since this is still ***uncharted territory*** with regard to the literature on MI, the chapter restricts imputation methods to different combinations of the current number of clusters at each level i.e. $4 \times 20 \times 100$.

Chapter 6

Simulation Study Part - 2

6.1 Varying Number of Clusters & Adjusting Confidence Intervals

The previous chapters discussed how to perform multilevel multiple imputation in three level datasets, the simulation study design in this thesis (section 4.2), data generating mechanisms (see section 4.2.1) and the implementation of extensions of MICE and JoMo for three level data (see sections 4.4 and 4.5). We discussed the performance of these extensions of MICE and JoMo when handling different proportions and mechanisms of missingness across different levels in the dataset. Finally, we also compared the performance of two approaches i.e. Passive Imputation and JAV approach for handling derived variables.

However, one might wonder if the performance of these methods are constrained by the fixed numbers of clusters defined at each levels i.e. $4 \times 20 \times 100$. To pose this as a question, *are the extended methods of multiple imputation effective across different values of clusters?*. Even more broadly, is multiple imputation in general robust to varying numbers of clusters at each level?

The structure of the simulation study is such that we have 4 individuals in each unit in the lowest level, who are nested within 20 individuals at level two, which is further nested within 100 individuals at the third (highest) level. This model was chosen as it closely represented the National Family Health Survey design, which is our illustrative example for the MICE (Ext.) and JoMo (Ext.) (see section 7.2). For the

purpose of this chapter, we interchange the cluster size at each level to understand how this may impact the estimates obtained from MI.

Austin and Leckie (2018) discuss the effect of varying the number of clusters and cluster sizes on statistical power of a test when using multilevel regression models. The objective of the study was to examine the effect of varying cluster sizes to detect a non-zero variance component. This study was extensive and simulated 836 scenarios across various cluster sizes. This chapter does not employ such an extensive simulation set-up and looks at varying the ratios of the existing cluster sizes across the three levels to study the potential impact this might have on the performance of MI.

Advancements within the field of multilevel multiple imputation are fairly recent and can be back tracked to only a few years ago (van Buuren, 2018; Robitzsch et al., 2019; Quartagno & Carpenter, 2016). Due to this, the effect of varying number of clusters is still an unexplored concept. The simulation results from the previous chapters provide us with the best leverage to explore this further.

6.2 Aims

As described above, in this chapter we compare the performance of the extended methods of MI i.e. MICE (ext.) and JoMo (ext.), across varying number of clusters at each level. We maintain the same performance measures as in the previous chapters (see Table 5.3.1 for definitions). We use absolute relative bias, mean square errors and probability coverage of intervals (or credible intervals) to measure the performance of the method for each scenario. Unlike the previous chapter, in this chapter, we adjust for bias in the coverage probabilities in confidence intervals. The bias in the intervals is calculated and then used to adjust the interval estimate obtained.

6.3 Method

The structure of the simulation study is similar to the previous chapter, as described in section 5.3. To recap, we simulate a complete dataset, introduce missing values to make in incomplete, apply MICE (ext.) and JoMo (ext.) to impute for the missing values, and perform standard statistical analysis on the imputed datasets and infer from the pooled estimates obtained.

In the previous chapter, we constructed a dataset with 4 units at level 1, 20 units at level 2 and 100 units at level 3. Thus, the ratio of the number of units at Level 1 : Level 2 : Level 3 is 1 : 5 : 25. We consider different permutations of these ratios. To make sensible inferences, not all permutations are considered. As the objective is exploratory, we restrict our methodology to include the following four scenarios:

1. **Scenario 1:** 4*100*20 ($L_1 : L_2 : L_3 = 1 : 25 : 5$)
2. **Scenario 2:** 20*4*100 ($L_1 : L_2 : L_3 = 5 : 1 : 25$)
3. **Scenario 3:** 100*20*4 ($L_1 : L_2 : L_3 = 25 : 5 : 1$)
4. **Scenario 4:** 4*20*100 ($L_1 : L_2 : L_3 = 1 : 5 : 25$)

Missing values are introduced into the dataset. To avoid making too many comparisons on the performance of MICE (ext.) and JoMo (ext.) across varying cluster sizes, and different mechanisms of missingness (see Tables 5.4.4 & 5.4.3 for this), we limit our simulation study to data that is Missing at Random (MAR) with 20% missingness observed at Level 1 and Level 2. This is equivalent to the Scenario 2 in chapter 5.

6.3.1 Results

Table 6.3.1 reports the performance of MICE (ext.) and JoMo (ext.) across all four scenarios listed above. At a glance, it is obvious that the number of clusters at each level has some effect on how MI performs. This is visible in the values of coverage of

Table 6.3.1: Performance of Multiple Imputation methods (ext.) across varying numbers of clusters at each level

Variables	**MICE (ext.)**			**JoMo (ext.)**		
	ARB	*MSE*	*Coverage (%)*	*ARB*	*MSE*	*Coverage (%)*
Scenario 1: 4*100*20 ($L_1 : L_2 : L_3 = 1 : 25 : 5$)						
Intercept	0.3287	0.9951	94.1%	0.3287	0.9951	91.1%
$X_{ijk} - \bar{X}_{jk}$	0.1348	0.1674	51%	0.1338	0.1503	97.4%
$\bar{X}_{jk} - \bar{X}_k$	0.1470	0.1935	52.3%	0.1347	0.1426	0%
$Z_{jk} - \bar{Z}_k$	0.3004	0.7304	13.8%	0.1418	0.1862	19%
W_k	0.2895	0.5382	93.5%	0.2896	0.5382	91.5%
Scenario 2: 20*4*100 ($L_1 : L_2 : L_3 = 5 : 1 : 25$)						
Intercept	0.2177	0.4612	100%	0.2176	0.4612	94.6%
$X_{ijk} - \bar{X}_{jk}$	0.1576	0.3207	0%	0.1335	0.1500	99%
$\bar{X}_{jk} - \bar{X}_k$	0.2489	0.5550	0%	0.1357	0.1412	37.2%
$Z_{jk} - \bar{Z}_k$	0.1933	0.4001	99.9%	0.1500	0.2569	97.4%
W_k	0.2839	0.5068	100%	0.2839	0.5068	94.5%
Scenario 3: 100*20*4 ($L_1 : L_2 : L_3 = 25 : 5 : 1$)						
Intercept	0.9639	10.0835	83.7%	0.9639	10.0836	75.1%
$X_{ijk} - \bar{X}_{jk}$	0.2365	0.4963	17.5%	0.1334	0.1499	91.3%
$\bar{X}_{jk} - \bar{X}_k$	0.4088	1.2591	10.2%	0.1448	0.1726	87.7%
$Z_{jk} - \bar{Z}_k$	0.2045	0.4185	72.7%	0.1637	0.3010	99.6%
W_k	0.3551	1.0255	85.1%	0.3557	1.0322	76.2%
Scenario 4: 4*20*100 ($L_1 : L_2 : L_3 = 1 : 5 : 25$)						
Intercept	0.2098	0.4333	100%	0.2158	0.4497	95.1%
$X_{ijk} - \bar{X}_{jk}$	0.1429	0.2082	0%	0.1340	0.1505	100%
$\bar{X}_{jk} - \bar{X}_k$	0.2037	0.4259	0%	0.1349	0.1422	0.01%
$Z_{jk} - \bar{Z}_k$	0.1487	0.2416	99.5%	0.1421	0.1825	4%
W_k	0.2846	0.5072	100%	0.2849	0.5068	94.3%

Note: *ARB = Absolute Relative Bias (see Table 5.3.1 for definition)*

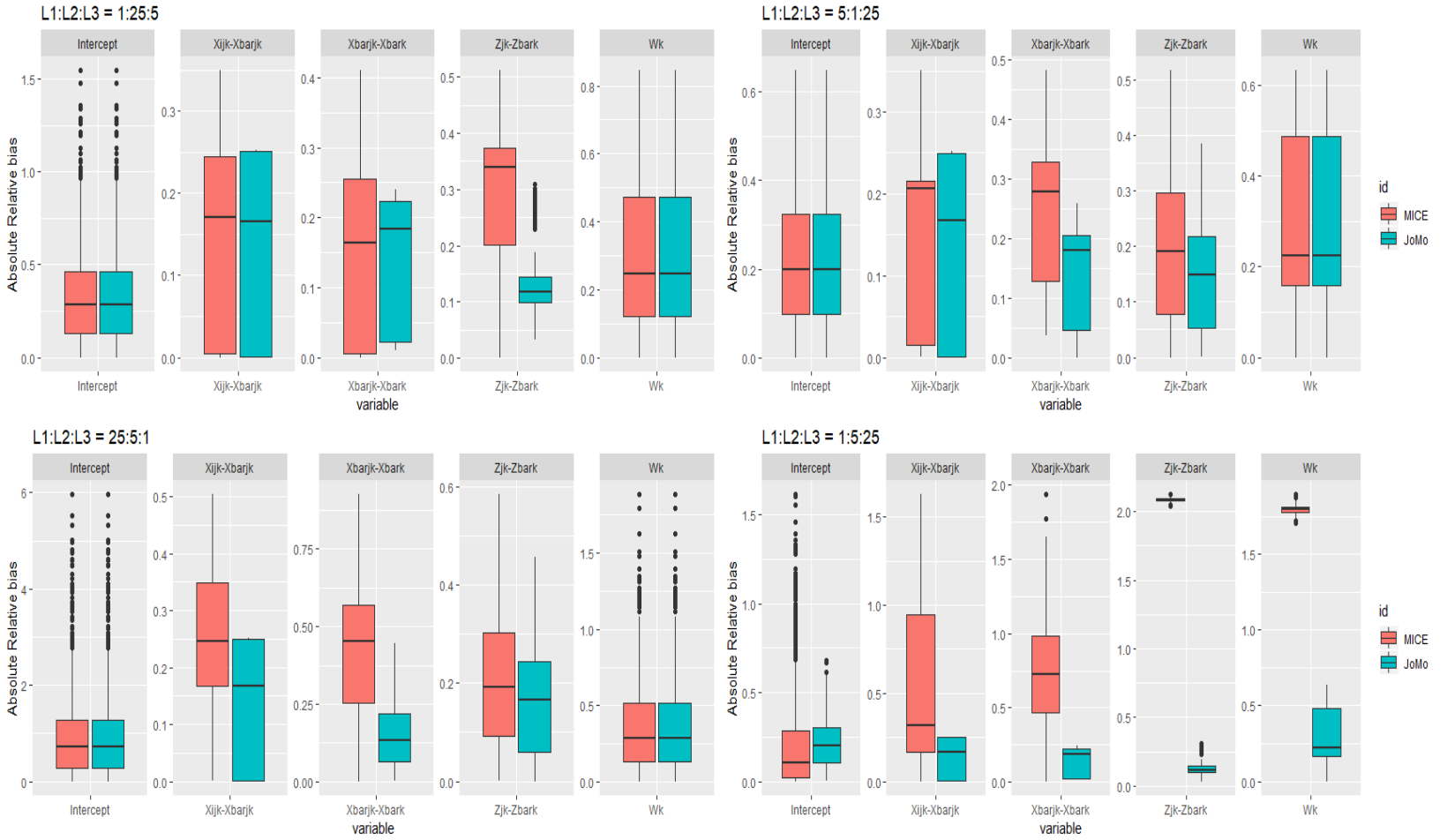


Figure 6.3.1: Distribution of Absolute Relative Bias of MI methods across different number of clusters from 1000 simulations

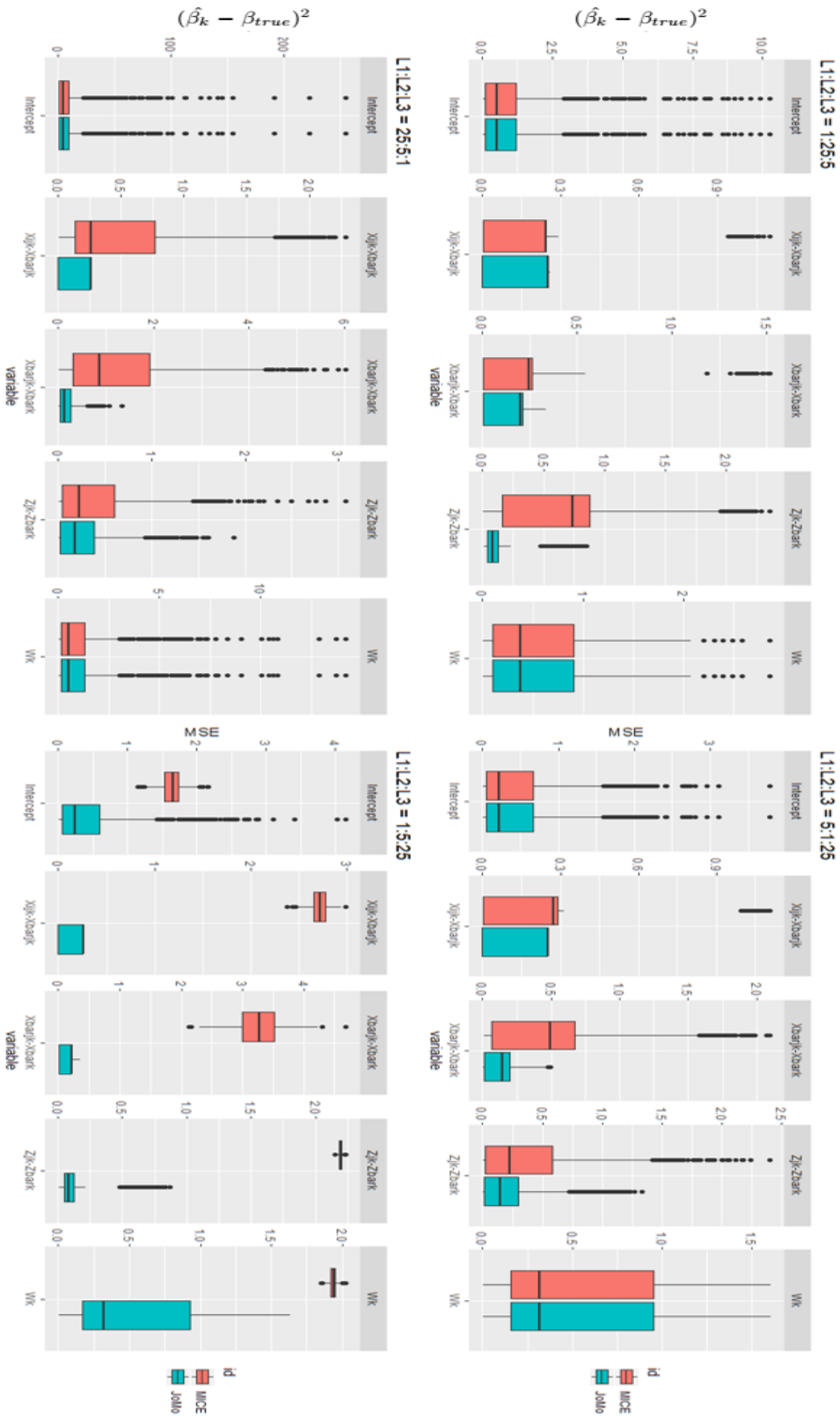


Figure 6.3.2: Distribution of $(\hat{\beta}_k - \beta_{true})^2$; (k indexing simulations) of MI methods across different number of clusters from 1000 simulations

confidence intervals, absolute relative bias and mean square error. We observe that across a variety in the number of clusters at each level, the extensions of MICE and JoMo show low values of bias and mean squared errors. Performance of JoMo (ext.) in Scenario 3 is worth noting as Table 6.3.1 shows low ARB and MSE values, and good probability of coverage of confidence intervals.

In comparison, our original cluster sizes of 4*20*100 i.e. $L_1 : L_2 : L_3$ = 1:5:25 (used for the simulation) show lower values of coverage. The reason for lower values of coverage has been discussed extensively in section 5.6.3. It is positive to observe lower values of absolute relative bias and MSE across the board (see Table 6.3.1). This suggests that the method MICE (ext.) is robust across varying numbers of clusters at each level (at least within this context). Figures 6.3.1 and 6.3.2 provide a deeper insight into the distribution of MSE and Bias across all four scenarios.

From Table 6.3.1, we see a huge variation in coverage of confidence intervals. This is extremely obvious for the MICE(ext.) in scenario 2: 20*4*100 ($L_1 : L_2 : L_3 = 5 : 1 : 25$); where the probability coverage of confidence intervals are at either 100% coverage or 0%. This was discussed in the Section 5.6.1, and we confirmed credibility of estimates obtained using credible intervals (see Figure 5.4.4). Given that coverage of confidence intervals varies across all the scenarios, it might be wise to check the credible intervals. We constructed a 95% Bayesian interval estimate by identifying the top 0.25% and bottom 0.25% of the estimates from the simulations. A Bayesian interval estimate is called a credible interval. Using much of the same notation as above, the definition of a credible interval for the unknown true value of θ is, for a given γ,

$$Pr(u(x) < \Theta < v(x)|X = x) = \gamma$$

Figure 6.3.3 gives the 95% credible interval for each scenario. The black triangles represent the true parameter values (see section 4.2.3 for values).

While Table 6.3.1 is useful in providing an overview of the absolute relative bias

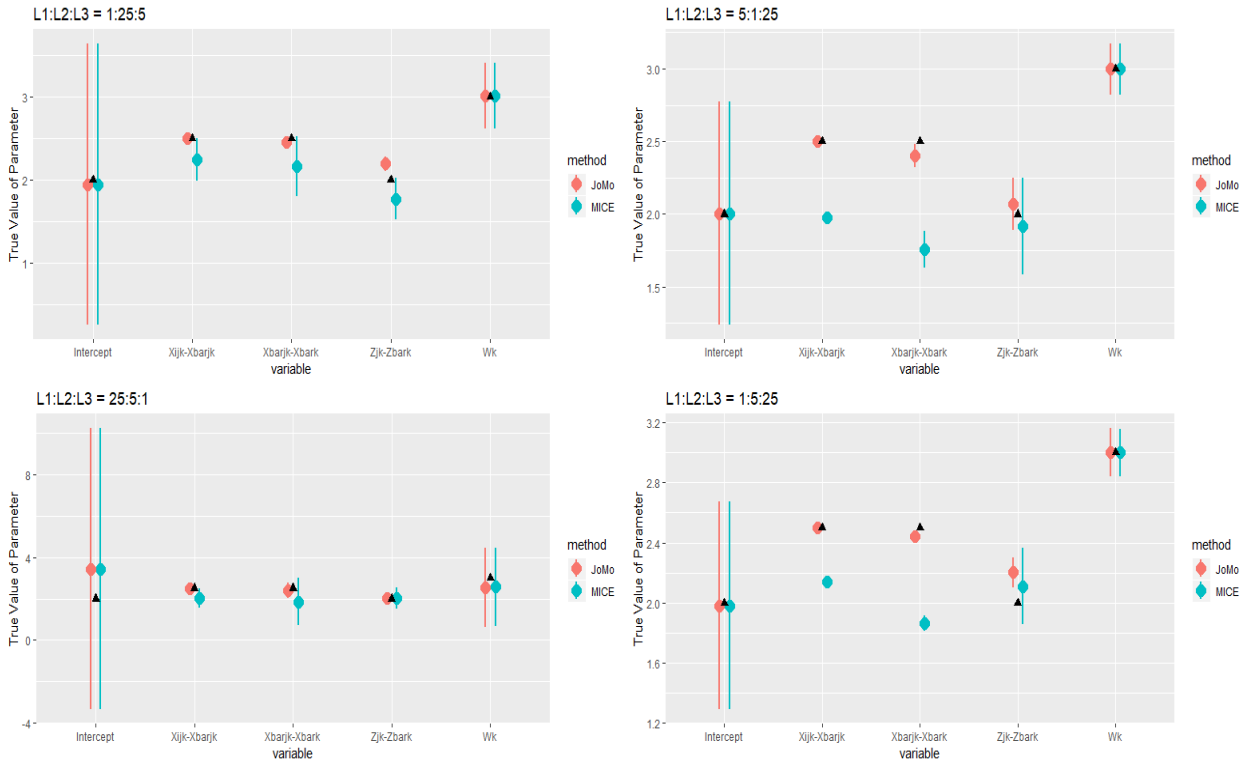


Figure 6.3.3: Credible Intervals obtained from 1000 simulations across all 4 scenarios of varying number of clusters. Black triangles represent the true parameter values.

and MSE observed across all four scenarios, a much clearer understanding of the distribution of these performance measures is obtained from Figures 6.3.1 and 6.3.2. We see that across the four scenarios, while ARB values are quite small, estimates obtained from JoMo (ext.) are smaller than those obtained from MICE (ext.). A similar observation can be made on comparing the credible intervals obtained across the four scenarios (see Figure 6.3.3), where we see that the credible intervals produced by the estimates obtained from JoMo (ext.) hover around the true parameter values (denoted by the black triangles) more than those calculated using MICE (ext.).

6.4 Adjustment of Coverage Probabilities of Confidence Intervals

Across both chapters 5 and 6, we have observed low probability coverage of confidence intervals for the estimates obtained from the MI methods. One of the most common causes for undercoverage is if bias $\neq 0$. Undercoverage due to bias should degrade as the number of observations increases (Morris et al., 2019). However, our results from the various scenarios suggests that while the bias in our results is small, our estimates obtained from the MI methods are not unbiased. Our interval estimates do not have the correct nominal coverage probabilities as expected (see Tables 5.4.4 and 6.3.1).

Menéndez, Fan, Garthwaite, and Sisson (2014) proposed a novel method to correct confidence intervals to have nominal coverage probabilities. This method does not make any distributional assumptions and the correction proposed in asymptotically unbiased. The theory behind the correction proposed is as follows:

If θ is the parameter of interest and $x \sim f(x|\theta)$, let $L(x)$ and $U(x)$ be estimators of the lower and upper limits of a $100(1-\alpha)\%$ level confidence interval. If $P(\theta < L(x)) \neq \frac{\alpha}{2}$, then the new estimator $L_c(x) = L(x) + \epsilon_{\alpha/2}$ where $\epsilon_{\alpha/2}$ is the $\alpha/2^{th}$ quantile of the distribution function of $G_{\{\theta - L(x)\}}$, so that

$$G_{\{\theta-L(x)\}}(\epsilon_{\alpha/2}) = \frac{\alpha}{2}$$

Thus the new estimator $L_c(x)$ will have the correct coverage probability

$$P(\theta < L_c(x)) = \alpha/2$$

This theorem proposed by Menéndez et al. (2014), is extended to adjust the upper limit of the confidence interval such that,

$$U_c(x) = U(x) + \epsilon_{1-\alpha/2}$$

where $\epsilon_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of the distribution function $G_{\{\theta-U(x)\}}$. Thus the new estimator $U_c(x)$ will have the correct coverage probability. The detailed derivations of this can be found in the paper Menéndez et al. (2014).

$$P(\theta > U_c(x)) = \alpha/2.$$

Correction Process as proposed by Menéndez et al. (2014)

1. Obtain the upper and lower limits of the $100(1-\alpha)\%$ CI for parameter θ.
2. Evaluate $\hat{\theta}$ and generate n independent datasets.
3. For each dataset, compute $L_c(x)$ and $U_c(x)$ for the $100(1-\alpha)\%$ CI.
4. Set the corrected upper and lower limits using the $(1-\alpha/2)^{th}$ quantile of the distribution functions $G_{\{\theta-U(x)\}}$ and the $\alpha/2^{th}$ quantile of the distribution function of $G_{\{\theta-L(x)\}}$.

6.4.1 Results

Table 6.4.2 provides the probability of coverage of both the adjusted and unadjusted confidence interval. The methods to adjust for bias as proposed by Menéndez et al. (2014) are robust to the varying number of clusters at each level and can be applied to adjust for bias in the confidence intervals produced by MI methods. We see that the 0% coverage of confidence intervals; which was possibly due to a result of the shift in the distribution of estimates obtained due to passive imputation of the derived variables, increases significantly to demonstrate a higher coverage of the parameter. The adjustment of the confidence intervals as demonstrated by the procedure in section 6.4, is not influenced by the method of multiple imputation applied or the proportions of missing values. Thus the process of bias correction in the confidence interval is independent of the missing data mechanisms and the method of imputation used. Menéndez et al (2014) comment on the robustness of the method, however, it is positive to see that this holds true in our simulation study as well.

Table 6.4.2: Comparing nominal coverage versus adjusted coverage of confidence interval

Variables	**MICE (ext.)**		**JoMo (ext.)**	
	Coverage	$Coverage_{adj}$	*Coverage*	$Coverage_{adj}$
Scenario 1: 4*100*20 ($L_1 : L_2 : L_3 = 1 : 25 : 5$)				
Intercept	94.1%	74.6%	91.1%	74.5%
$X_{ijk} - \bar{X}_{jk}$	51%	60.7%	97.4%	75%
$\bar{X}_{jk} - \bar{X}_k$	52.3%	60.6%	0%	74.9%
$Z_{jk} - \bar{Z}_k$	13.8%	73.8%	19%	75%
W_k	93.5%	74.8%	91.5%	74.7%
Scenario 2: 20*4*100 ($L_1 : L_2 : L_3 = 5 : 1 : 25$)				
Intercept	100%	75%	94.6%	75%
$X_{ijk} - \bar{X}_{jk}$	0%	75%	99%	75%
$\bar{X}_{jk} - \bar{X}_k$	0%	75%	37.2%	75%
$Z_{jk} - \bar{Z}_k$	99.9%	75%	97.4%	75%
W_k	100%	75%	94.5%	74.8%
Scenario 3: 100*20*4 ($L_1 : L_2 : L_3 = 25 : 5 : 1$)				
Intercept	83.7%	73.4%	75.1%	70%
$X_{ijk} - \bar{X}_{jk}$	17.5%	15.9%	91.3%	75%
$\bar{X}_{jk} - \bar{X}_k$	10.2%%	15.9%	87.7%	74.7%
$Z_{jk} - \bar{Z}_k$	72.7%	73.7%	99.6%	75%
W_k	85.1%	73.2%	76.2%	69.4%
Scenario 4: 4*20*100 ($L_1 : L_2 : L_3 = 1 : 5 : 25$)				
Intercept	100%	75%	95.1%	74.8%
$X_{ijk} - \bar{X}_{jk}$	0%	75%	100%	75%
$\bar{X}_{jk} - \bar{X}_k$	0%	75%	0.01%	74.9%
$Z_{jk} - \bar{Z}_k$	99.5%	75%	4%	75%
W_k	100%	75%	94.3%	75%

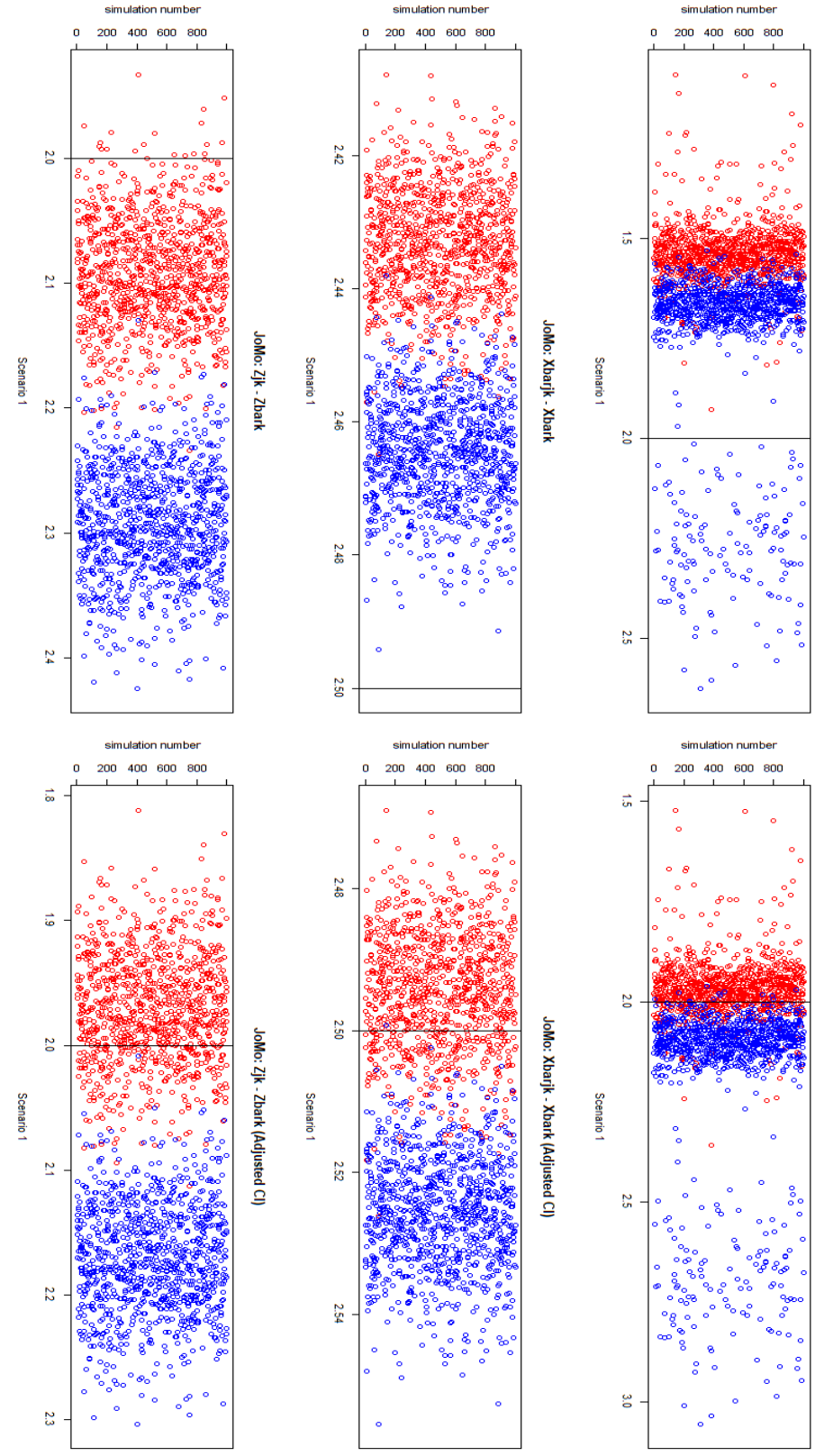


Figure 6.4.4: Highlighting Coverage of CIs for parameter estimates after adjusting for bias (Scenario 1)

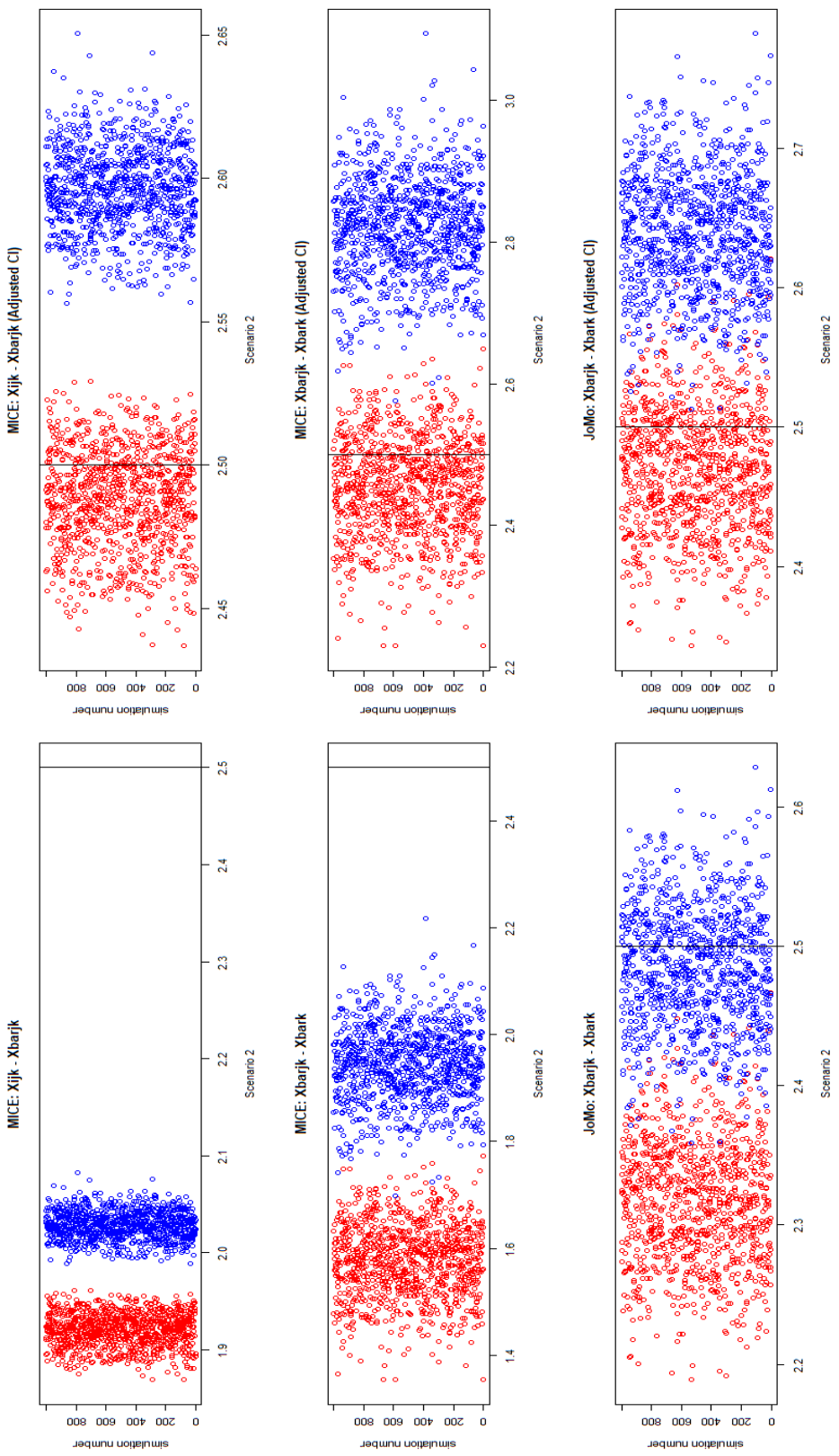


Figure 6.4.5: Highlighting Coverage of CIs for parameter estimates after adjusting for bias (Scenario 2)

Figures 6.4.4 and 6.4.5 further illustrate the dramatic increase in the probability coverage of the confidence intervals after adjustment for bias. The blue dots in the figure represent the upper limits of the confidence intervals and the red dots represent the lower limits of the confidence interval. The black line shows the true values of the parameters (see Table 4.2.3). The graphs show the confidence intervals for 1000 simulations before and after adjustment of the confidence intervals. It is positive to observe the adjusted confidence intervals capturing the true value of the parameter post adjustment as illustrated for Scenarios 1 and 2.

6.5 Summary

In this chapter, we highlighted the performance of MI methods across varying number of clusters at each level in the simulated datasets. We concluded that the extended methods of MICE and JoMo continue to perform well with low values of absolute relative bias and MSE. We see that lower values of bias and MSE are reported as the number of units at the first and second level increases. In addition, we have also corrected the low coverage values observed in Table 6.3.1. We observe an increase in the coverage probabilities after adjustment of confidence intervals.While this adjustment method was developed in the frequentist framework, the approach is general and makes minimal assumptions, allowing for wide applicability. The method provides a reliable approach to adjust approximately obtained credible intervals in dynamic settings.

Up to this point, we observed and compared the performance of extended methods of MI in a controlled setting using simulated datasets. However, no experiment is complete until tested in the real world. To this end, in the next chapter we test the performance of extensions of MI, in handling missing values in national surveys. In addition, we also study the impact of multiple imputation in analysis and reiterate the need to account for data structures in data analysis.

Chapter 7

Application of Multilevel Modelling and Multiple Imputation

7.1 Description

This thesis falls at the intersection of two areas of focus in public health statistics - **missing data** and **multilevel data**. The two topics frequently occur across medical research as investigators often have to deal with missing values as a consequence of conducting a study beyond the laboratory (a clean and controlled setting), and most study designs are interested in the role of contextual factors or clusters in affecting the outcome variable. The core of the thesis up till now has focused on extending the current methods of multiple imputation (see Chapter 4) and has used simulations as a proof of concept (see Chapter 5). However, no experiment is successful until it has been tested in uncontrolled settings.

In this chapter, we highlight the importance of each key core aspect the thesis, i.e. using multiple imputation effectively, accounting for high dimensional datasets in analysis, and the intersection that is practical demonstration of the impact of using multiple imputation in accommodating missing data in multilevel models.

7.2 Case Study 1: Application of three level multiple imputation in national surveys

7.2.1 Introduction

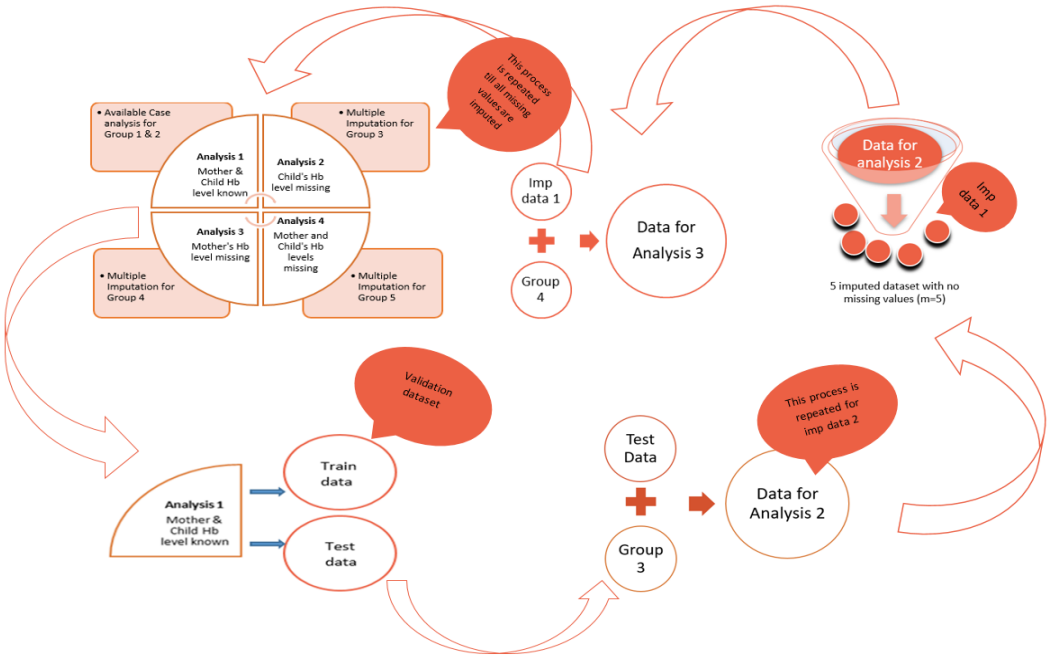
In this first case study, we use the fourth National Family Health Survey (NFHS-4) data of India, to implement our extensions of multiple imputation. In this study, we investigate the prevalence of maternal and child anaemia in Northern India. The study is naturally hierarchical with children nested within mothers who are further nested within households. Anaemia is a multifaceted issue with individual, biological, and environmental risk factors. The prevalence of anaemia in India is on the rise, with a high concentration of maternal and child anaemia cases in the North of India.

Due to the impact of household, maternal and individual risk factors on child anaemia levels, we have included predictor variables across these three levels in the dataset. The ability to implement multilevel multiple imputation is vital in such cases. Presence of missing data in any of these levels can result in underestimation of the actual prevalence and in turn influence implementation of national and state policies targeting this issue. This study aims to identify maternal and household level predictors of anemia in children in India. This study is novel in its approach in imputing for missing data in the third level (households) in the NFHS survey for India.

Estimating Effect of Maternal and Household Risk Factors Influencing Anaemia among Children under 5 years Using Multiple Imputation

Nidhi Menon, Alice Richardson, Hwan Jin Yoon

1 Graphical Abstract



1.1 Anaemia in India

According to the World Health Organization (WHO), anaemia is one of the ten most serious health problems in the world. Globally, no country is on the track to achieving the 2025 targets to improve maternal, infant and young child nutrition. Nutritional deficiency can lead to severe anaemia in the population. In children, anaemia may contribute to poor growth, stunting, developmental cognitive delay and increased susceptibility to infections (Nguyen et al., 2018). Anaemia has also been known to impair work productivity, and increase adverse pregnancy outcomes, low birth weight, small for gestational age infants. South East Asia and Africa continue to have the highest prevalence of anaemia and account for about 85% of the burden, affecting mainly women and children. Prevalence of anaemia among women in India has seen little improvement, witnessing a decline from 55% in 2005 - 2006 to just 53% in 2015-2016 (IIPS & ICF, 2017). In India, 70% of adolescent girls are anaemic and over 50% are below normal body mass index (BMI). With a prevalence higher than 40 %, WHO considers anaemia in children 6 - 59 months as a severe public health problem.

Globally 1/3rd of the population lack household level access to a safe and reliable supply of drinking water, and safe & acceptable forms of sanitation (WHO, 2019). Sub-Saharan Africa and South Asia account for greater deficits in access to safe water and sanitation. Water, sanitation, and hygiene (WASH) practices, in addition to safe water supply and adequate sanitation accompanied by proper hygiene education can significantly reduce risk of infections, and improve health outcomes (Kothari et al., 2019).

India is undergoing an unprecedented process of economic, demographic and social transformation spanning over the last few decades. India plays a significant role in global health due to both its geographical and population size. Recent years have seen extraordinary progress towards both improving health and spreading awareness through government initiatives and intervention of international agencies. Despite these developments, the practice of public health in India has been non-uniform and has witnessed several hurdles. The country has the largest number of stunted children in world i.e. 46.8 million approx (Haddad et al., 2015).

India is undergoing an unprecedented process of economic, demographic and social transformation spanning over the last few decades. India plays a significant role in global health due to both its geographical and population size. Recent years have seen extraordinary progress towards both improving health and spreading awareness through government initiatives and intervention of international agencies. Despite these developments, the practice of public health in India has been non-uniform and has witnessed several hurdles. Although urban areas have better access to safe drinking water compared to rural areas, the proportion of the urban population with access to safe services is falling as government investments fail to keep up with the growing urban population. High disparities in hand washing practices between urban and rural India are still observed. Over 60% of the world's open defecation happens in India. At 732 million, India tops list on the number of people without access to toilets. Indian states with poor access to sanitation report high incidence of diarrhoeal diseases. Uttar Pradesh, Bihar, Madhya Pradesh, Assam and Chhattisgarh had the highest prevalence of both anaemia and diarrhoeal diseases(IIPS & ICF, 2017).

The causes of anaemia are multifaceted, spanning between biological, social and environmental. Water, sanitation and hygiene features at various levels in this framework with varying degrees of proximity to the outcome. Existing literature shows significant progress in understanding the relationship between WASH and nutrition, with WASH having been listed as one of the key interventions to improve nutrition (Cumming & Cairncross, 2016). To further strengthen this evidence, we evaluate anaemia prevalence in the context of WASH in India.

The paper builds on the current body of literature and will provide analysis from the nationally representative National Family Health Survey. It is also apparent that providing reliable estimates of anthropometric indicators is important for monitoring the country's progress towards reducing health inequalities.

The objective of this study is to illustrate the uses of multiple imputation when accounting for any missingness that the survey brings to the analysis. A high degree of clustering in the proportion of anaemic women are observed in the northern states of India (IIPS & ICF, 2017; Swaminathan et al., 2019). Our study aims on identifying socio-demographic, individual and

household factors associated with maternal and child anaemia in North India using multiple imputation. The study uses a novel approach of combining mother's and child's missing data in the imputation process, and combining this imputed data with the original dataset to obtain reliable estimates.

2 Methods

2.1 Data

The NFHS-4 dataset has a naturally hierarchical structure where information is captured for all children (6-59 months) of interviewed women within the households selected for interview. Each household has one mother with an average of 3 children per mother. The structure of the data includes children nested within interviewed women for each household that are further nested within primary sampling units. Information is captured for children, women and household. To evaluate the missing values in the data, and the relationship between maternal and environmental factors on the child anaemia; we classified children into 5 groups as shown in Figure 1. Group 1 and Group 2 can be considered as complete data as information on the anaemic status for both mothers and their children in known. The two groups were not combined as this gives us the opportunity to compare the profiles of combination of non- anaemic mother and children with their counterparts. Groups 3 and 4 are scenarios where information on anaemic status has been captured at only one level. Group 5 has missing values observed at both levels. Here we consider the children to be at level 1, mothers to be at level 2 and households at level 3.

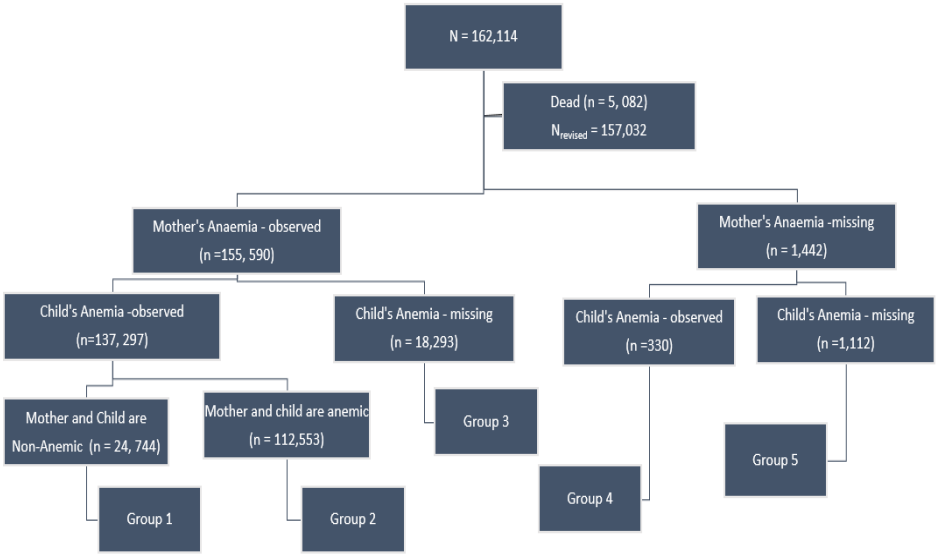


Figure 1: Individuals grouped according to patterns of anaemia observed and missing across mother and child

This approach of classification based on patterns of missing values is inspired from a longitudinal study focused on identifying relationship between work-family conflict and mental health (Nguyen et al., 2018). Our study extends this approach to a clustered hierarchical study design using the NFHS survey data. Such hierarchical structures are commonly referred to as multilevel data structures. Multilevel linear regression models were fit for each of the 5 scenarios. Child's haemoglobin levels (g/dl) - a continuous measure, is the outcome variable. This survey is particularly important as it is the first Indian community wide survey to provide estimates for blood glucose and blood pressure levels in the general population.

The blood specimens were collected from eligible women (15-49 years), men (15-54 years) and children (6-59 months) using a drop of blood from a finger prick (or heel prick for children. A portable HemoCue Hb 201+ analyser was used to perform on-site haemoglobin (Hb) analysis. Respondents found to have severe anaemia were referred to a health facility for treatment. Severe anaemia is defined as Hb levels less than 9grams/decilitre(g/dl) for pregnant women, and less than 7g/dl for all others.

Other covariates considered in the study can be categorised into 3 groups: Child, maternal and household, and Primary Sampling Unit (PSU). Some covariates measured for the child include age, birth weight, and recent occurrence of diarrhoea. Birth weight is very important to predict future growth (Bharati, Pal, Bandyopadhyay, Bhakta, & Chakraborty, 2011). Mother's education level, BMI, anaemia status, age and number of antenatal visits are some maternal factors that can affect child malnutrition. We have already established WASH as a key intervention to improve nutrition. We use classifications of Household Condition Index as defined by Bharati et al.(2011), to reflect the household's condition with respect to sanitation, water and hygiene. We derived 3 variables using this classification. Each of the variables pertaining to toilet facilities, water and family size are scored according to the quality of facility available for the household.

- Toilet facilities are scored as 0: no facility; 1: shared/public pit toilet; 2: shared/public flush toilet or own pit toilet; 3: own flush toilet.
- Source of drinking water is scored as 0: not piped water e.g. surface water, river, dam etc.; 1: public tap, hand-pump or well; 2: private pipelines, pump, or well in own yard/plot.
- Family size is scored as 0: 11 or more members; 1: eight to ten members, 2: five to seven members; 4: less than 5 members

We first check if there are any differences between the groups with respect to demographic characteristics (see Table 1). We hypothesise that the missing groups may have a different demography and can help provide some insight into the varying attributes of each of these groups. We performed a step wise multiple imputation process. The first imputed dataset at each step was used as prior information to be used to build the imputation model in the succeeding analysis. Each iteration becomes more comprehensive compared to available case analysis as we include more observations with partially observed variables, reflecting the more complex and intricate health scenario in India. Analysis 2 uses information captured in Groups 1 & 2, drawing in the observed values for haemoglobin levels measured for children to impute for missing Hb values in Group 3. The first imputed dataset at each step was used as prior information to be used to build the imputation model in the succeeding analysis. Analysis 4 uses MI to impute for missing values for maternal haemoglobin levels captured in the dataset. Analysis 5 used MI to impute for missing values in haemoglobin levels for both mother and child.

Table 1: Demographic characteristics by missing data groups

Variable	Group 1 (n = 24, 744)	Group 2 (n = 112, 553)	Group 3 (n = 18, 293)	Group 4 (n = 330)	Group 5 (n = 1,112)	p-value
Child's Characteristics measured						
Age (years)	2.633 (1.24)	2.192 (1.29)	0.554 (1.21)	2.275 (1.29)	1.860 (1.47)	< 0.001
Birth weight (kg)	2.861 (0.581)	2.802 (0.59)	2.813 (0.58)	2.802 (0.62)	2.803 (0.63)	< 0.001
Hb level (g/dl)	11.936 (0.768)	10.534 (1.41)	NA	10.006 (1.74)	NA	< 0.001
Diarrhoea						< 0.001
(no)	22,186 (89.66%)	99,671 (88.55%)	15,429 (84.34%)	281 (85.15%)	974 (87.59%)	
(yes, last 2 weeks)	2,537 (10.25%)	12,821 (11.39%)	2,705 (14.79%)	47 (14.24%)	131 (11.78%)	
(dont' know)	21 (0.08%)	61 (0.05%)	159 (0.87%)	2 (0.61%)	7 (0.63%)	
Maternal and Household Characteristics measured						
Age (years)	28.155 (5.28)	27.642 (5.49)	25.436 (5.23)	27.363 (5.69)	26.956 (5.09)	< 0.001
BMI	21.824 (5.03)	21.050 (4.51)	21.173 (3.99)	21.838 (5.05)	21.422 (3.99)	< 0.001
Hb level (g/dl)	12.888 (0.85)	11.031 (1.52)	11.224 (1.656)	NA	NA	Outcome
Education						< 0.001
(None)	8,789 (35.51%)	45, 287 (40.24%)	6, 140 (33.56%)	109 (33.03%)	414 (37.23%)	
(Primary)	3,104 (12.54%)	14,407 (12.80%)	2,214 (12.10 %)	43 (13.03%)	131 (11.78%)	
(Secondary)	9,523 (38.49%)	41,668 (37.02%)	7,556 (41.31%)	139 (42.12%)	400 (35.97%)	
(Higher)	3,331 (13.46%)	11,191 (9.94%)	2,383 (13.03%)	39 (11.82%)	167 (15.02%)	
Wealth Index						< 0.001
(Poorest)	7,030 (28.41%)	36,482 (32.41%)	5,529 (30.22%)	94 (28.48%)	330 (29.68%)	
(Poorer)	5,280 (21.34%)	23,420 (20.81%)	3,743 (20.46%)	67 (20.30%)	183 (16.46%)	
(middle)	4,014 (16.22%)	18,359 (16.31%)	3,094 (16.91%)	55 (16.67%)	171 (15.38 %)	
(richer)	3,680 (14.87%)	16,254 (14.44%)	2,648 (14.48%)	47 (14.24 %)	179 (16.10%)	
(richest)	4,740 (19.16%)	18,038 (16.03 %)	3,279 (17.92%)	67 (20.30%)	249 (22.39%)	
Antenatal Visits						< 0.001
(0 visits)	4,406 (24.35%)	20,538 (24.71%)	3,655 (21.92%)	74 (28.79%)	258 (28.93%)	
(<10 visits)	13,023 (1.96%)	60.173 (72.39%)	12,497 (74.96%)	180 (70.04%)	603 (67.68%)	
(10-30 visits)	608 (3.36%)	2,158 (2.60%)	458 (2.75%)	0	24 (2.69%)	
(Don't know)	61 (0.34%)	249 (0.30%)	62 (0.37%)	3 (1.17%)	6 (0.67%)	
Household Structure						< 0.001
Nuclear Family	8,092 (32.70%)	37,909 (33.68%)	4,400 (24.05%)	88 (26.67%)	321 (28.87%)	
Non-nuclear family	16,652 (67.30%)	74,644 (66.32%)	13,893 (75.95%)	242 (73.33%)	791 (71.13%)	
HCI toilet^{+}						
(No facilty)	12,278 (49.62%)	59,059 (52.47%)	9,288 (50.77%)	160 (48.48%)	541 (48.65%)	
(public pit toilet)	187 (0.76%)	1,050 (0.93%)	150 (0.82%)	2 (0.61%)	14 (1.26%)	0.562
(public flush toilet)	1,803 (7.29%)	8,459 (7.52%)	1,179 (6.45%)	18 (5.45%)	71 (6.38%)	< 0.001
(own flush toilet)	10,476 (42.34%)	43,985 (39.08%)	7,676 (41.96%)	150 (45.45%)	486 (43.71%)	< 0.001
HCI water^{+}						< 0.001
(not piped water)	298 (1.20%)	1,370 (1.22%)	224 (1.22%)	8 (2.42%)	17 (1.53%)	
(public pipelines)	19,300 (78%)	90,132 (80.08%)	14,480 (79.16%)	247 (74.85%)	836 (75.18%)	
(private pipelines)	5,146 (20.80%)	21,051 (18.70%)	3,589 (19.62%)	75 (22.73%)	259 (23.29%)	
HCI family size^{+}						
($\geq$ 11 members)	3,110 (12.57%)	14,171 (12.59%)	2,672 (14.61%)	71 (21.52%)	179 (16.10%)	
(8-10 members)	6,082 (24.58%)	26,950 (23.94%)	4,874 (26.64%)	93 (28.18%)	267 (24.01%)	0.090
(5-7 members)	11,598 (46.87%)	53,567 (47.69%)	8,098 (44.27%)	123 (37.27%)	487 (43.79%)	0.001
($<$ 5 members)	3,954 (15.98%)	17,754 (15.77%)	2,649 (14.48%)	43 (13.03%)	179 (16.10%)	0.018
PSU Characteristics measured						
State						< 0.001
(Bihar)	5,586 (22.58%)	28,041 (24.91%)	4,402 (24.06%)	66 (20%)	204 (18.35%)	
(Haryana)	1,298 (5.225%)	10,339 (9.19%)	1, 612 (8.81%)	41 (12.42%)	107 (9.62%)	
(Jharkhand)	2,344 (9.47%)	15,356(13.64%)	2,352(12.86%)	42 (12.73%)	203 (18.26%)	
(Punjab)	1,732 (7%)	6,459 (5.74%)	1,041 (5.69%)	9 (2.73%)	62 (5.58%)	
(Uttar Pradesh)	11727 (47.39%)	42,211 (41.06%)	7,635(41.74%)	127 (32.48%)	430 (38.67%)	
(Uttarakhand)	2,057 (8.31%)	6,147 (5.46%)	1,251 (6.84%)	45 (13.64%)	106(9.53%)	
Place of Residence						< 0.001
(Urban)	5,586 (22.58%)	23,931 (21.30%)	3,829 (20.93%)	71 (28.55%)	318 (28.55%)	
(Rural)	19,158 (77.42%)	88,582 (78.70%)	14,464 (79.07%)	259 (78.48%)	796 (71.45%)	

Analysis 1 : group 1+ group 2 ; **Analysis 2 :** 25% (group 1+ 2) + Group 3; **Analysis 3 :** Imputed dataset from Analysis 2 + group 4; **Analysis 4 :** Imputed dataset from Analysis 3 + group 5
Continuous variables are summarised using Mean(S.D) and categorical variables are summarised using frequency (%)
$^{+}$ refer to the HCI classification explained in section **??**

2.2 Analysis and Handling Missing Data

Risk factors were identified using information from previous literature and variables pertaining to WASH. Bivariate analysis was performed to determine impact of selected risk factors on outcome variable (child's Hb levels). It can be noted that the large sample size may be highly influential in any resulting statistical significance. Table 1 reports corresponding p-values obtained from the bivariate analysis. The dataset was then split into 2 parts: 25% of the dataset was used to train the model and 75% of the dataset was used to test the model obtained.

2.3 Imputation Procedure

Missing values in the dataset were imputed using Multiple Imputation using Chained Equations (MICE). All analysis was performed using the package (mice) in R (version 3.4.0). We performed a sequential method of imputation based on the groups of missing data patterns. The target variables for the imputation models are haemoglobin levels (g/dl) captured for both mother and child, birth weight of the child, number of antenatal visits, mother's BMI, presence of water at household, and household type as defined by NFHS. Therefore, information was imputed at each level i.e. household, mother and child. The target variables here are the variables with missing values in the dataset, and should not be confused with the outcome variable for the analysis model.

2.4 Post Imputation Procedure

After completing the cycle of imputations, we now have no missing data in any variable. We imputed for missing Haemoglobin (Hb) levels for both mother and children. Based on these imputed values, we can classify each mother and child as non-anaemic or anaemic. Individuals with Hb level $<$11 g/dl are considered non anaemic and those with Hb levels $<$ 11g/dl are categorised as anaemic. It is important to note that the Hb values were imputed on a continuous scale and then dichotomously categorised (see Table 2). We observe that all individuals are now reclassified and no missing values are observed in the dataset for both maternal and child Hb levels.

This approach of classification based on patterns of missing values is inspired from a longitudinal study focused on identifying relationship between

Table 2: Classification of all Mothers and Children after Imputation of Haemoglobin levels

Anaemia Status	Mother (Non Anaemic)	Mother (Anaemic)
Child (Non Anaemic)	12,201 (Model 1)	21,620 (Model 2)
Child (Anaemic)	5,463 (Model 3)	14,850(Model 4)

work-family conflict and mental health (Nguyen et al., 2018). Our study extends this approach to a clustered hierarchical study design using the NFHS survey data. Such hierarchical structures are commonly referred to as multilevel data structures. Multilevel linear regression models were fit for each of the 5 scenarios. Child's haemoglobin levels *(g/dl)* - a continuous measure, is the outcome variable. We first check if there are any differences between the groups with respect to demographic characteristics (see Table 1). We hypothesise that the missing groups may have a different demography and can help provide some insight into the varying attributes of each of these groups.
We performed a step wise multiple imputation process. The first imputed dataset at each step was used as prior information to be used to build the imputation model in the succeeding analysis. Each iteration becomes more comprehensive compared to available case analysis as we include more observations with partially observed variables, reflecting the more complex and intricate health scenario in India. Analysis 2 uses information captured in Groups 1& 2, drawing in the observed values for haemoglobin levels measured for children to impute for missing Hb values in Group 3. The first imputed dataset at each step was used as prior information to be used to build the imputation model in the succeeding analysis . Analysis 4 uses MI to impute for missing values for maternal haemoglobin levels captured in the dataset. Analysis 5 used MI to impute for missing values in haemoglobin levels for both mother and child.

In order to compare how well our imputations worked, we compare these classifications to validation data (75% of original dataset). We construct a confusion matrix to show the performance of the imputation process in classifying mothers and children as anaemic and non-anaemic. A confusion matrix is a technique for summarizing the performance of a classification

Table 3: Measures for binary classification obtained to assess classification of all mother's and children based on Hb levels after imputation of missing data

Statistics	Child's Classification	Maternal Classification
Accuracy (95% CI)	0.9844 (0.983, 0.985)	0.9809 (0.979,0.982)
Sensitivity	0.9991	0.9979
Specificity	0.9540	0.9707
Positive Predictive Value	0.9782	0.9534
Negative Predictive Value	0.99881	0.9987
Detection Rate	0.6731	0.3744
Balanced Accuracy	0.9766	0.9843

algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating confusion matrix statistics (see Table 3) provides a better idea of what your classification model is getting right and the of errors it is making. The key concept of confusion matrix is that it calculates the number of correct & incorrect predictions which is further summarized with the number of count values and breakdown in each class of outcome.

3 Results

3.1 Demographic characteristics by missing data group

The results from Table 1 indicate many demographic differences between the groups, when comparing across maternal, child and household characteristics. We cannot simply categorise a group as advantaged or disadvantaged using the mean values of these variables. The distribution of maternal age across the 5 groups is very similar. The average age of mothers is between 26-27 years. However, the most crucial observation in this table would be the difference in the child's age between the groups. We observe that groups where Hb levels were missing on an average had children under 2 years of age compared to the other groups (Group 3: 0.554 (1.21), Group 5: 1.860

(1.47)). This seems plausible as mothers of children under the age of 2 years may not consent to the heel prick required by the HemoCue test.

The average BMI of these women is in the healthy range (18 - 25 kg/m^2) across the groups, however on comparing the distribution, we observe that group 1 has less that 10% of women classify as underweight, while under around 25% of the women in groups 2 - 5 are classified as underweight. We also observe several outliers (positive), indicating more women with BMI ≥ 30 in groups 1,2 and 3. This number is considerably lower when plotting BMI values for women in groups 4 and 5. We also observe an increase in the proportion of children who had diarrhoea in the last two weeks, in groups 3,4 and 5.

The highest number of study participants were from Uttar Pradesh and Bihar. We observe that across the groups, 30% of mothers have received no education, and nearly 15% are educated up to the primary level. There is a stark difference in access to toilet facilities. Around 50% of households in the states considered in the study have no toilet facilities, while more than 40% of the remaining households own their own flush toilets. We also observe that most households are classified as Poorer or Poorest, and less than 20% of households have under 5 members. This indicates large family structure (non-nuclear families) which is characteristic of the nation. However, this also points towards strong competition for resources and unequal access to facilities by all members of the household. Similarly, a large proportion of households depend on public water sources that are shared among multiple households. Public water sources include public tap, hand-pump or a shared well.

3.2 Estimates obtained after Imputation of Missing Data

Before, we proceed with post imputation analysis, it is important to check the imputations. This is done by plotting the distribution of the observed and the imputed data for every variable imputed. We observe that the distributions of the imputed values overlaps with the distribution of the observed data (see Supplementary Figure 1a). This is one of the most popular method of diagnostic check and is referred to as internal checks since the imputed

data is being assessed with respect to the data in hand. These checks to help the user identify potential issues with the imputation model. These are key observations to consider in the thesis and by extension in any multilevel MI analysis, as the most challenging task in MI is the specification of the imputation model. Any misspecification in the model would result in biased estimates. These diagnostic measures are used to draw inference on the validity of the imputation model. We observe that the distributions of the imputed values overlaps with the distribution of the observed data (see see Supplementary Figure 1a) indicating proper imputations of variables. In addition, we also confirm convergence of the imputations by inspecting the trace lines post imputations. This has been plotted for variables, Maternal and child's Hb levels, mother's BMI and child's birth weight in see Supplementary Figure 1b.

The results of the multilevel linear regression model of maternal anaemia on the child's anaemia measured after adjusting for covariates listed in the Supplementary Table 1. In general we see that effects are positive, indicating a positive impact of increased maternal haemoglobin values on the child's haemoglobin levels. The results of the regression analysis are differentiated across the four models based on maternal and child anaemia classification (see Table 2). Imputations were performed for Haemoglobin levels (g/dl) that was observed as continuous variable, and then dichotomised into anaemic or non-anaemic based on the cut-off value of 11 g/dl. Based on the imputations performed, we reclassify every mother and child into anaemic or non-anaemic. We compare classification rates after imputations with the validation dataset. We see an accuracy of 0.9844 (0.983, 0.985) observed when classifying children as anaemic or non-anaemic and an accuracy of 0.9809 (0.979, 0.982) when classifying mothers.

It is important to note that all children in models 1 & 3 are non-anaemic and children in models 2 & 4 are anaemic. From Supplementary Table 1, we observe that across all groups, child's age, birth weight and maternal haemoglobin levels has a positive impact on the child's haemoglobin levels. A similar observation was made for *number of antenatal visits*. In groups where mother's were not anaemic, we observed a positive impact of having 10-30 antenatal visits on the child's Haemoglobin levels.

In accordance to the public health emphasis on WASH, we observe that

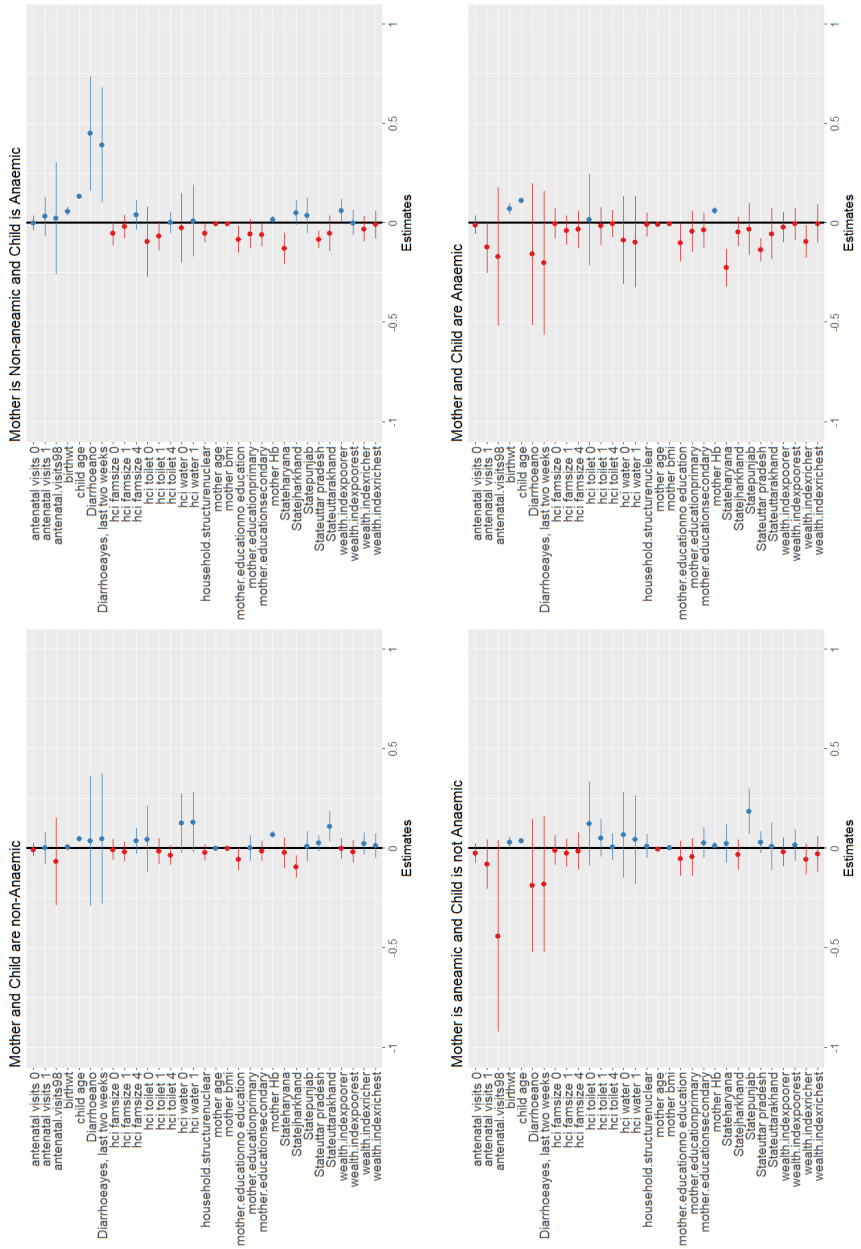
access to better sanitation and hygiene shows an improvement in the Hb levels across the 4 models compared to the households with no toilet facilities. Similar results were observed when comparing the impact of reduced household sizes on child's anaemia status compared to households with ≥ 11 members. We see that children belonging to states Jharkhand, Punjab and Haryana reported lower Hb levels.

4 Discussion

There is an urgent need to emphasise on the nutritional status of Indian women and children in the coming generations, not only to implement policy to improve their current health status but also to preserve the economic development of the nation in the near future. Differentials in under nutrition among children has been studied in relation to maternal nutritional status, by considering factors pertaining to child, mother and household levels.

The obstacle of missing data, was overcome by constructing proper imputation models that accounted for factors measured at every level in the dataset and included all variables included in the analysis model. The four stage imputation process that segregated the sample based on maternal and child haemoglobin levels observed proved to be accurate (98%) and showed high sensitivity and specificity values. While this strategy was inspired and adopted from a social science study looking and parental employment and health (Nguyen et al., 2018); it is positive to note that the stage wise analysis method is robust in imputing alternate health parameters in large demographic datasets. However, we deviate from their proposed methodology by reclassifying the mothers and children into anaemic and non-anaemic categories, and analysing them based on their imputed values. Another difference in the two studies is the multilevel structure of the dataset that allows us to see the role of individual,maternal and household variables on the outcome.

Figure 2: Comparison of estimates obtained across all four analysis groups



We observe that child's age and maternal haemoglobin levels have a significant impact on the child's haemoglobin levels. A key finding in the study is the positive impact of increased antenatal visits (10-30 visits) on the child's haemoglobin levels. This was particularly observed in cases where mothers were classified as non-anaemic (Model 1 & Model 2). This is a particularly important finding in the context of North India, as we see that less than 5% of women made more than 10 antenatal visits (see Table 1). We observe that children growing in nuclear household structures report lower Haemoglobin levels compared to joint families. There is evidence suggesting that in non-western societies, women and children's health and nutrition is influenced by multiple members of the family (Aubel, 2012). This includes the nutrition/health of pregnant, breastfeeding women and young children.

Another interesting find was that, with the exception of the group with non-anaemic mother and child, richer and richest households showed lower Haemoglobin levels compared to the middle income households. This finding contradicts some of the literature (Rammohan, Awofeso, & Robitaille, 2011), which identifies women and children from higher wealth quintiles to have reduced odds of iron-deficiency anaemia. However, we believe that this finding further cements our theory of anaemia being a multifaceted issue that cannot be boxed into a purely socio-economic condition affecting only the poorer households in the Indian context.

Mother's education has consistently been identified as a critical factor in determining the health of the child. From Supplementary Table 1, we see that children of women with no, primary and secondary education all reported lower Haemoglobin levels that children of women higher education. Despite, the several government initiatives targeted at improving women literacy in India, we still observe around 15% of women have completed higher (or tertiary) education. It is important to note that while 35% of women have completed their secondary education, we still have about the same proportion of women who have received no education. It is extremely important to note that these women have a mean age of 28 years, and are eligible to actively participate in the workforce and contribute to the household finances. Currently, Indian women constitute only 25% of India's labour force and contribute towards only 18% of the countries GDP. One of the major barriers, has been reported as the burden of household work on Indian girls. Woetzel et al. (2015) reported that housework accounts for 85% of the time women

in India spend on unpaid household work. The impact of women's literacy on child health has been established for decades, including assertions that the ability to send a child to school depends on a certain level of household income. The Right to Education Act implemented in 2010, mandates that every child between the ages of 6 to 16 receives free and compulsory education. Improving female literacy would prove beneficial in improving the child's nutrition status, while also contributing significantly to the countries economic growth.

Finally, water, sanitation and hygiene (WASH) have been established as essential in early childhood development. Prevalence of under nutrition in children in the study can be associated with poor WASH practices. Access to proper toilets continue to remain a challenge, with a disappointingly large proportion of households reporting No toilet facilities (approximately 50%). While the rest of the world eliminates open defecation, this practice continues to prevail in India. According to UNICEF, faecal contamination and poor sanitation is a leading cause of childhood mortality disease and under nutrition. The government's ambitious sanitation campaign aimed at making India defecation free, with building more than 110 million toilets around the country. Despite this, there continues to remain a challenge in changing the attitudes of many especially in the rural areas about open defecation, who consider it cleaner than having a toilet inside their home (Coffey et al., 2014).

We observe that households with access to public or private pipelines as source of household water, show a positive impact on child's haemoglobin levels compared to households using surface water sources. It is positive to note that about 20% of households have access to private pipelines, while most households (approximately 80%) use public pipelines. Water quality is always a public health concern especially in India as the country lacks the water treatment facilities that can handle water contamination by industries, and open defecation that is prevalent across the country. Thus proper household water treatment practices becomes increasingly important to achieve safe supply of drinking water (Li, Liu, & BeLue, 2018). This study relies on the type of water supply instead of the measures taken to treat the water for consumption, it is hard to establish the exact impact of water quality on child's anaemia status. The practice of boiling water prior to consumption is prevalent across the country and may be a factor that contributes to the positive effect observed for water consumed from public pipelines. As observed

in the literature, drinking unclean water is a major cause for diarrhoea in children. The possibility of proper water treatment approaches, may have also contributed to the low prevalence of diarrhoea (15%) observe in the samples. In contrast, households with large family sizes has been shown to have a positive impact of the child's nutrition. This has been discussed previously and is in alignment with our findings on the positive impact of non-nuclear households on child's health.

5 Conclusion

Although there are many studies addressing factors associated with anaemia in the Indian population, there is still limited literature on handling missing values while accounting for the complex sampling strategy in analysis. This study is among the first few illustrations on imputing for missing values in a three level data structure. The study focuses on how the treatment of missing data will systematically exclude mothers and infants. A number of social and economic disadvantages are associated with young mothers in north India, including low levels of education, limited peer networks and restricted mobility. The consequential exclusion of a large proportion of women and children using sub-standard methods of analysis of missing data further disadvantages these mothers and children rendering their health status invisible and underestimated. The study also shows an increased need for public heath intervention in north India. We sought to capture the full range of maternal and child anaemia patterns and illustrated the need to focus on how integral missing data treatment can be.

References

Aubel, J. (2012). The role and influence of grandmothers on child nutrition: culturally designated advisors and caregivers. *Maternal & child nutrition*, *8*(1), 19–35.

Bharati, P., Pal, M., Bandyopadhyay, M., Bhakta, A., & Chakraborty, S. (2011). Prevalence and causes of low birth weight in india. *Malaysian journal of nutrition*, *17*(3).

Coffey, D., Gupta, A., Hathi, P., Khurana, N., Spears, D., Srivastav, N., & Vyas, S. (2014). Revealed preference for open defecation. *Economic & Political Weekly*, *49*(38), 43.

Cumming, O., & Cairncross, S. (2016). Can water, sanitation and hygiene help eliminate stunting? current evidence and policy implications. *Maternal & child nutrition*, *12*, 91–105.

Haddad, L., et al. (2015). Actions and accountability to advance nutrition and sustainable development. *Globalizations*.

IIPS, & ICF. (2017). *National family health survey (nfhs-4), 2015–16.*

Kothari, M. T., Coile, A., Huestis, A., Pullum, T., Garrett, D., & Engmann, C. (2019). Exploring associations between water, sanitation, and anemia through 47 nationally representative demographic and health surveys. *Annals of the New York Academy of Sciences*, *1450*(1), 249.

Li, W., Liu, E., & BeLue, R. (2018). Household water treatment and the nutritional status of primary-aged children in india: findings from the india human development survey. *Globalization and health*, *14*(1), 37.

Nguyen, C. D., Strazdins, L., Nicholson, J. M., & Cooklin, A. R. (2018). Impact of missing data strategies in studies of parental employment and health: Missing items, missing waves, and missing mothers. *Social Science & Medicine*, *209*, 160–168.

Rammohan, A., Awofeso, N., & Robitaille, M.-C. (2011). Addressing female iron-deficiency anaemia in india: is vegetarianism the major obstacle? *ISRN Public Health*, *2012*.

Swaminathan, A., Kim, R., Xu, Y., Blossom, J. C., Venkatraman, R., Kumar, A., ... others (2019). Burden of child malnutrition in india: A view from parliamentary constituencies. *Economic & Political Weekly*, *54*(2).

Woetzel, J., et al. (2015). *The power of parity: How advancing women's equality can add 12 trillion to global growth.*

Specifying the Imputation Model

The Imputation Model for birthweight is specified as follows. Here, the target variable, $birthwt_{ijk}$ is the birth weight off the i^{th} child of the j^{th} mother in the k^{th} household; i = 1, 2, ...n, j = 1, 2.. J and k = 1,2...K.

$$birthwt_{ijk} = \beta_{0jk} + \beta_1 Age_{ijk} + \beta_2 Hb_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01} mothersage_{jk} + \gamma_{02} Edu_{jk} + \gamma_{03} antenatalvisits_{jk} + \gamma_{04} maternalbmi_{jk} + r_{0jk}$$

$$\gamma_{0k} = \delta_{000} + \delta_{001} housestructure_k + \delta_{002} state_k + \delta_{001} houseBPL_k + u_{00k}.$$

Here the covariates are as follows, Age_{ijk} : Age of child, Hb_{ijk}: Hb level of the child, $mothersage_{jk}$: Mother's age, Edu_{jk} : Highest education level for mother, $antenatalvisits_{jk}$: Total number of antenatal visits during pregnancy, $maternalbmi_{jk}$: Mother's BMI levels, $housestructure_k$: Household structure (nuclear or not), $state_k$: State, $houseBPL_k$: Household has a Below Poverty Line (BPL) card.

Imputation model for Antenatal visits are specified as follows:

$$antenatalvisits_{jk} = \gamma_{00k} + \gamma_{01} mothersage_{jk} + \gamma_{02} Edu_{jk} + \gamma_{03} anganwadi_{jk} + \gamma_{04} maternalhb_{jk} + r_{0jk}$$

$$\gamma_{00k} = \delta_{000} + \delta_{001} wealthindex_k + \delta_{002} state_k + \delta_{001} houseBPL_k + u_{00k}.$$

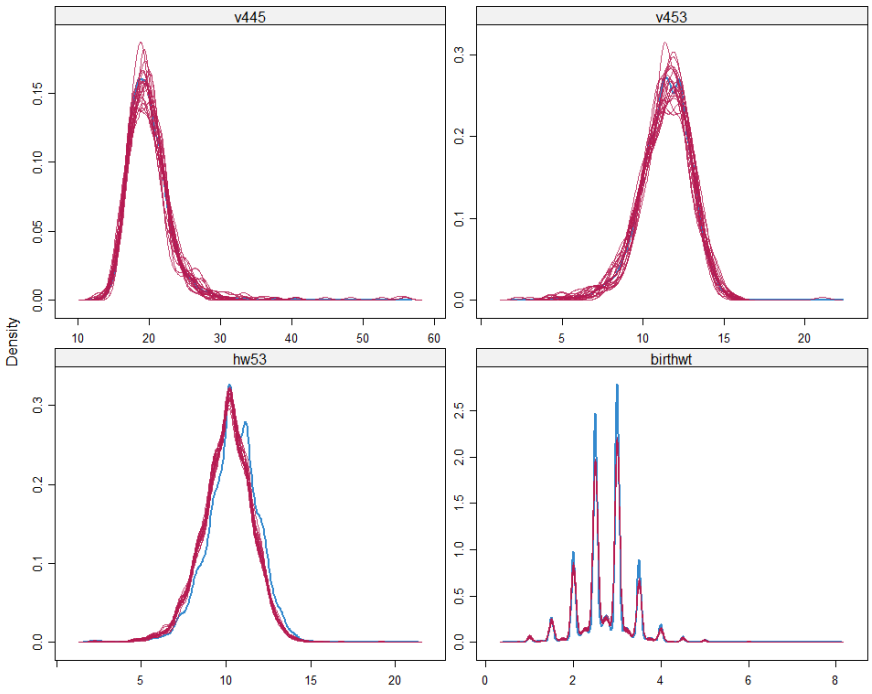
Where the covariates are: $mothersage_{jk}$: Mother's age, Edu_{jk} : Highest education level for mother, $anganwadi_{jk}$: Met with an anganwadi worker in last 3months, $wealthindex_k$: Household wealth index

Table 1: Parity per mother

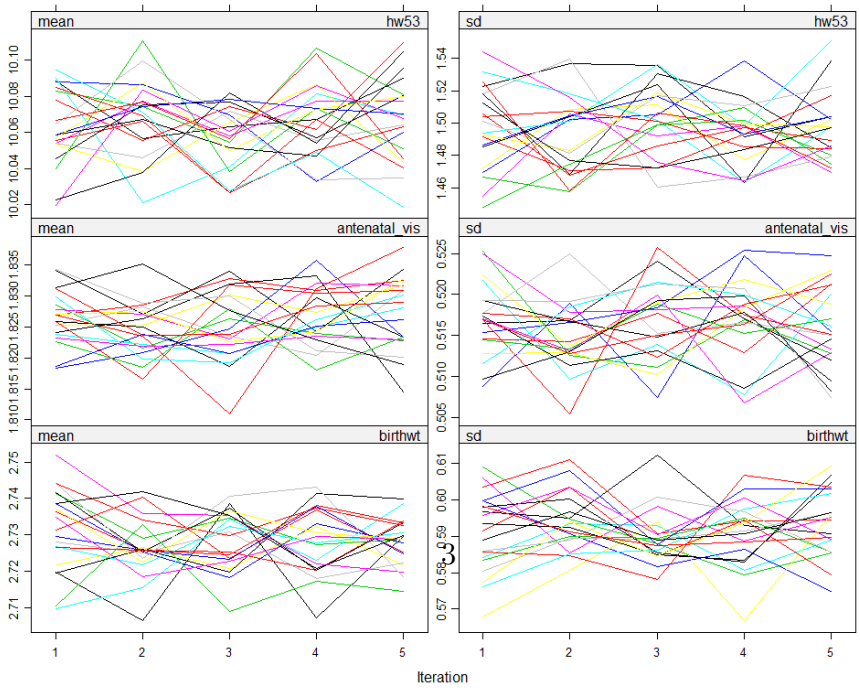
Number of Children per mother in the dataset	Frequency
1	21,446
2	85,823
3	49,417
4	14,807
5	4140
6	1023
7	247
8	63
9	21
10	7
11	6
12	1
14	1

Figure 1: Checking Imputations

(a) Distribution of imputed versus observed values



(b) convergence plots



v445: mother's BMI **v453**: mother's Hb level
hw53: child's Hb level **birthwt**: child's birthweight

Table 2: Results of Analysis by Maternal and Child Anaemia Reclassification Groups

Variable	Model 1 $\hat{\beta}$ (SE)	Model 2 $\hat{\beta}$ (SE)	Model 3 $\hat{\beta}$ (SE)	Model 4 $\hat{\beta}$ (SE)
Intercept	10.81 (0.204)*	8.840 (0.219)*	11.62 (0.285)*	9.039 (0.255)*
Child's Characteristics measured				
Age (years)	0.0500 (0.005)*	0.1348 (0.006)*	0.0382 (0.007)*	0.113 (0.007)*
Birth weight (kg)	0.0065 (0.010)*	0.0602 (0.011)*	0.0315 (0.143)*	0.0735 (0.014)*
Diarrhoea				
(no)	0.0431 (0.166)	0.457 (0.146)*	-0.186 (0.170)*	-0.155 (0.182)
(yes, last 2 weeks)	0.0562 (0.168)	0.0392 (0.147)*	-0.179 (0.173)	-0.198 (0.184)
Maternal and Household Characteristics measured				
Age (years)	0.0075 (0.002)	-0.0042 (0.035)*	-0.0031 (0.002)	-0.0048 (0.002)*
BMI	-0.0003 (0.002)	-0.0028 (0.002)	0.0039 (0.003)	-0.0029 (0.003)
Hb levels	0.0072 (0.007)*	0.0177 (0.009)	0.01307 (0.009)	0.0623 (0.009)*
Education				
(No Education)	-0.0542 (0.029)*	-0.0822 (0.035)*	-0.0492 (0.044)	-0.0976 (0.049)*
(Primary)	0.0017 (0.033)	-0.0531 (0.038) *	-0.0413 (0.048)	-0.0412 (0.052)
(Secondary)	-0.0124 (0.026)	-0.0573 (0.031)	0.0294 (0.039)	-0.0347 (0.045)
Antenatal Visits				
(<10 visits)	-0.0048 (0.016)	0.0077 (0.019)	-0.0231 (0.024)	-0.0087 (0.023)
(10-30 visits)	0.0026 (0.042)	0.0346 (0.050)	-0.0790 (0.064)	-0.1192 (0.067)
(Don't know)	-0.0666 (0.112)	0.0252 (0.144)	-0.4382 (0.244)	-0.1682 (0.178)
Household Structure				
(Nuclear family)	-0.0154 (0.022)	-0.0501 (0.024)*	0.0125 (0.030)	-0.0077 (0.034)
HCI toilet$^+$				
(public pit toilet)	0.0459 (0.085)	-0.09320 (0.090)	0.1250 (0.106)	0.0176 (0.118)
(public flush toilet)	-0.0129 (0.034)	-0.0631 (0.037)	0.0532 (0.048)	-0.0141 (0.048)
(own flush toilet)	-0.0333 (0.024)	0.0030 (0.027)	0.0086 (0.034)	-0.0015 (0.035)
HCI water$^+$				
(Public pipeline)	0.1270 (0.076)	-0.0237 (0.089)	0.0691 (0.110)	-0.0851 (0.112)
(Private pipeline)	0.1308 (0.076)	0.0127 (0.091)	0.0443 (0.114)	-0.0939 (0.116)
HCI family size$^+$				
(8-10 members)	-0.0047 (0.027)	-0.0524 (0.031)	-0.0086 (0.037)	-0.0021 (0.039)
(5-7 members)	-0.0162 (0.025)	-0.0168 (0.029)	-0.0221 (0.036)	-0.0356 (0.037)
(< 5 members)	-0.0333 (0.024)	0.0417 (0.037)	-0.0139 (0.047)	-0.0318 (0.048)
Wealth Index				
(Poorer)	-0.0004 (0.026)	0.0618 (0.030)*	-0.0166 (0.037)	-0.0195 (0.038)
(Poorest)	-0.0155 (0.028)	0.0018 (0.032)	0.0184 (0.040)	-0.0048 (0.041)
(richer)	0.0243 (0.028)	-0.0258 (0.033)	-0.0527 (0.039)	-0.0935 (0.042)*
(richest)	0.0151 (0.031)	-0.0073 (0.036)	-0.0254 (0.046)	-0.0033 (0.042)
PSU Characteristics measured				
State				
(Haryana)	-0.0212 (0.039)*	-0.1261 (0.041)*	0.0255 (0.049)	-0.226 (0.048)*
(Jharkhand)	-0.0912 (0.029)	0.0518 (0.031)	-0.0308 (0.038)	-0.0427(0.037)
(Punjab)	0.0099 (0.038)	0.0380 (0.046)	0.1867 (0.058)*	-0.0307 (0.067)
(Uttar Pradesh)	0.0271 (0.019)	-0.0822 (0.029)*	0.0316 (0.028)	-0.1342 (0.029)*
(Uttarakhand)	0.1102 (0.039)*	-0.0505 (0.046)	0.0112 (0.061)	-0.0534 (0.066)
Random Effects	Variance			
psu:hhid:id	0.2633	0.7167	0.2262	0.8445
hhid:id	0.1098	0.1216	0.1036	0.1837
psu	0.0198	0.0298	0.0199	0.0316
Residual	0.1165	0.2256	0.1208	0.2101

Model 1 : Mother and Child are non- anaemic ; **Model 2 :** Mother is non anaemic and Child is anaemic ; **Model 3 :** Mother is anaemic and Child is non anaemic ; **Model 4 :** Mother and Child are anaemic

* p-value < 0.05

$^+$ refer to the HCI classification explained in section **??**

7.3 Case Study 2: Impact of accounting for data structure in national surveys

7.3.1 Introduction

In this case study, we demonstrate the need to account for the data structure in analysis. This case study aims to identify the risk profile for HIV infection using an appropriate modeling strategy and highlight the consequences of ignoring the structure of the data. It offers a methodological guide towards an applied approach to the identification of future risk and the need to customize intervention to address HIV infection in the Indian population.

This is done using the third national family health survey for India (2005-2006), to identify risk factors for HIV. This study demonstrates the implications of proceeding without accounting for the multilevel structure of the data and ignoring the sampling technique. The study revealed certain insights into HIV risk factors in India. Higher risk was observed among households of higher wealth index. This observation contradicts the popular belief that poverty drives HIV epidemics as it leads people to high risk behaviours. Primary Sampling Units covered by Anganwadis explained to be a protective factor for HIV reinstating the importance of Anganwadis in controlling the spread of HIV in India. Results also showed a higher risk of HIV among married individuals among men, and all the women tested as HIV positive were married. This shows an increase in the risk of individuals getting infected through their spouses.

Several advantages of using multilevel modelling over the conventional multiple linear/logistic regressions are observed. The biggest advantage of proceeding with multilevel analysis is its capacity to produce reliable estimates. Failing to acknowledge the hierarchical structure of the data would result in over estimating the standard error; which in turn would end in wrong inferences. This paper appears in the Biostatistics and Epidemiology Journal and was published in collaboration with the National Institute of Mental Health and Neuro Sciences (NIMHANS), Bangalore, India.

BIOSTATISTICS & EPIDEMIOLOGY
2019, VOL. 3, NO. 1, 126–139
https://doi.org/10.1080/24709360.2019.1671096

Taylor & Francis
Taylor & Francis Group

Effect of modeling a multilevel structure on the Indian population to identify the factors influencing HIV infection

Nidhi Menon [a,b], Binukumar Bhaskarapillai [b] and Alice Richardson [a]

[a]National Centre for Epidemiology & Population Health, Australian National University, Canberra, Australia; [b]Department of Biostatistics, National Institute of Mental Health and Neuro Sciences (NIMHANS), Bangalore, India

ABSTRACT

Many studies have addressed the factors associated with HIV in the Indian population. Some of these studies have used sampling weights for the risk estimation of factors associated with HIV, but few studies have adjusted for the multilevel structure of survey data. The National Family Health Survey 3 collected data across India between 2005 and 2006. 38,715 females and 66,212 males with complete information were analyzed. To account for the correlations within clusters, a three-level model was employed. Bivariate and multivariable mixed effect logistic regression analysis were performed to identify factors associated with HIV. Intracluster correlation coefficients were used to assess the relatedness of each pair of variables within clusters. Variables pertaining to no knowledge of contraceptive methods, age at first marriage, wealth index and noncoverage of PSUs by Anganwadis were significant risk factors for HIV when the multileveled model was used for analysis. This study has identified the risk profile for HIV infection using an appropriate modeling strategy and has highlighted the consequences of ignoring the structure of the data. It offers a methodological guide towards an applied approach to the identification of future risk and the need to customize intervention to address HIV infection in the Indian population.

ARTICLE HISTORY
Received 11 April 2018
Accepted 8 September 2019

KEYWORDS
HIV infection; logistic regression; multilevel modeling; multistage sampling

1. Introduction

India stands second in the list of most populated countries in the world with a population of over 1.2 billion. Of this number, the number of individuals estimated to be living with Human Immunodeficiency Virus (HIV) is around 2.5 million between the ages of 15 and 49 years [1]. India also has the third largest number of people living with HIV in countries in the world [2]. The National Family Health Survey (NFHS)-3, a large scale survey conducted using a representative sample of households throughout India, estimated this country's adult HIV prevalence to be approximately 0.36% [1]. This figure is small compared to most other middle-income countries, yet translates to a population estimate of 2.1 million people living with HIV in India. The prevalence of HIV in India varies geographically and high prevalence has been reported from the states of Andhra Pradesh, Karnataka,

CONTACT Binukumar Bhaskarapillai binukumarb@gmail.com Associate Professor, Department of Biostatistics, National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru, Karnataka, India – 560029

Maharashtra and Tamil Nadu that accounts for 53% of all HIV infections. However, government initiatives such as Link Worker Scheme and Red Ribbon Express have succeeded in increasing awareness on methods to reduce risk of contracting HIV [2]. Further, an increase in the political motivation and commitment to prevention and control of HIV has led to various interventions and scaled-up prevention strategies under the National AIDS Control Program in India [2].

Sexual transmission of HIV is reported as the most predominant method by which the virus has spread across the country [3]. Studies have addressed the factors associated with HIV in the Indian population [4–13]. These studies identified various factors influencing HIV infection including lack of formal education [11], unprotected sexual intercourse [9–11], multiple sexual partner [4,12], intravenous drug use [10,13], transfusion of contaminated blood and blood products [10], presence or history of ulcerative STDs [10–12], irregular use of condoms [4,10,11], and current or previous genital ulcer or genital discharge [4,9,11]. In addition, another study that assessed the relation between STDs and HIV, reported a greater risk in rural subjects than in urban subjects [12]. For the prevention and control of HIV, Godbole proposed a multi-disciplinary combined approach of early detection, treatment of STDs, promotion of condoms, blood safety and drug de-addiction programs in order to increase awareness that might result in behavioral change among vulnerable populations in India [3].

Many of these studies referenced above have analyzed the survey data using national sampling weights [2,14]. The NFHS surveys require the use of sample weights to ensure sample representativeness. '*A weight is an inflation (or representative) factor applied to each unit in relation with the overall probability of selection and interview for each case in a sample*' [15]. Sampling weights are constructed so that each unit (respondent or household) can be inflated to represent other individuals or households in India. The NFHS dataset has different weights depending on the outcome and the unit of analysis i.e. households, women or children, men or HIV test result.

Analysis of survey data requires that researchers account for both sampling strategies employed in the survey by accounting for the survey weights as well as the nested structure of the data. Hence, this article demonstrates the impact of multilevel models compared to single-level models (not accounting for the hierarchical structure of data) in identifying the factors influencing the HIV epidemic in the Indian population using NFHS data. For the risk estimation of factors associated with HIV, the studies that have adjusted for the multilevel structure of survey data during analysis are negligible [16,17].

Multilevel modeling is a statistical technique that extends ordinary regression analysis to the situation where the data are hierarchical and accounts for that complex structure and dependencies of the observations at different levels of sampling by including multiple variance components, one at each level of the structure [18]. In India, national household surveys are conducted by considering divisions within the states and urban/rural specification; and primary sampling units (clusters in the NFHS survey). Thus, the sampling design is inherently clustered and hierarchical in nature. The sampling technique employed in NFHS-3 varies between urban and rural areas. In urban areas, wards were selected as the first level units with probability proportional to size (PPS) sampling from which census enumeration blocks (CEB) were selected at random and then households were selected from these CEBs using simple random sampling. In the case of rural areas, the primary sampling units (PSUs) were villages, from which households were selected at random. This

clearly implies a 3 level data structure with individuals nested within households, followed by households nested within PSUs. This structure further indicates that units within the same higher level unit are not independent. For example, we expect information captured for individuals living in the same household to be much strongly related to each other. This could be due to their Socioeconomic status (SES), education, religion, etc. In such cases, invoking the independence assumption leads to invalid inferences as the precision of the effects is often overestimated as variability between individuals in the same household is wrongly attributed to variability between individuals. The standard errors of any estimated parameters from a survey are calculated by allowing for survey weights along with the sampling strategy. Multilevel model based estimators of standard errors have the potential to account for hierarchical nature of the data and provide better estimates.

2. Setting

In response to the urgent need to have nationally-representative data on HIV prevalence and comprehensive information on knowledge and attitudes about HIV/AIDS, high-risk sexual behavior, and practices related to HIV testing in India, it was proposed to incorporate these issues in NFHS-3. It was planned to provide HIV prevalence levels for the population of women ages 15–49 years and men aged 15–54 years at the national level and for each of the six high HIV prevalence states as identified by National AIDS Control Organization (NACO), namely, Andhra Pradesh, Karnataka, Maharashtra, Manipur, Nagaland, and Tamil Nadu. Considering the large sample size in Uttar Pradesh it was further decided to provide estimates of HIV prevalence for the state of Uttar Pradesh also. All women aged 15–49 years and all men aged 15–54 years living in all sample households in those states were eligible for HIV testing. In the remaining 22 states, HIV testing was conducted in only a subsample of six households per enumeration area, and all women aged 15–49 and all men aged 15–54 in those sample households were eligible for HIV testing. NFHS-3 is the third in a series of national surveys and the first large scale nationwide survey in India to use Dried Blood Spots (DBS) technique for HIV testing. DBS samples were collected on filter paper using the finger prick technique in the field. These were then tested in one or more central laboratories. Individuals were classified as HIV positive/ HIV negative or indeterminate on the basis of the result of the DBS test. For this study data were obtained from the Demographic and Health Survey (DHS) website: www.dhsprogram.com

This data consists of a nationally representative sample of 1,09,041 households. There was a total of 1,24,385 women aged between 15 and 49 years, and 74,369 men aged between 15 and 54 years used for this study. Among all the women and men interviewed nationwide, 1,05,657 (39,257 women and 66,400 men) were tested for HIV. However, in the multilevel regression analysis to follow, a total of 38,715 females and 66,212 males with complete information were included. Survey weights relevant to the prevalence of HIV were used to account for the sampling strategy employed in the survey. In accordance to the standard NFHS guidelines, the HIV sample weight variable used in the analysis was (HIV05).

Socio-demographic factors such as gender, age, and education, coverage of ward/village by an Anganwadi, residence (urban/rural), marital status and household wealth index were available. Individual characteristics such as haemoglobin level, HIV test result, the number

of living children, ever had a blood transfusion, and history of sexually transmitted infections (STI)/ genital sore/ discharge in the last 12 months was also collected. Behavioral variables such as awareness about methods to reduce risk of acquired immunodeficiency syndrome (AIDS), ever heard of STI/ AIDS, the number of sexual partners, age at first marriage and paid for sex in the last 12 months were also included in the analysis.

In order to obtain a numerical measure of the awareness of individuals, the variable '*Awareness of methods to reduce the risk of AIDS*' was created. The variable was generated by adding the results from three awareness related variables namely, '*Reduce the chance of getting AIDS by not having sex at all*,' '*Reduce the chance of getting AIDS by always using condoms during sex*' and '*Reduce the chance of getting AIDS by having only 1 sexual partner*.' The values were recoded as '0' denotes 'Knows no method,' '1' denotes 'Knows one method' and '2' denotes 'Knows more than one method.' The outcome variable in the study, the HIV blood test result (positive/negative), is measured at the individual level.

3. Methods

In order to account for cluster sampling strategy employed in data collection; which if ignored would overestimate the standard deviation due to the effect of cluster variance; a multilevel modeling approach to the analysis was employed. The first level consists of the PSUs i.e. wards in urban areas and villages in rural areas. The second level is the households within each PSU. Individuals within each household make the final level in the data structure. Information is captured at each level in the data. A three-level survey design was declared for the dataset and a multilevel model was constructed with PSUs at level 1, households at level 2 and individuals at level 3. All statistical analysis was done using Stata (version 14 MP, Stata Corp, Texas, USA). A three-level survey design was declared for the dataset. This was done using the survey set command- '***svyset.***' The variable (V001) was used to identify the primary sampling unit – the first level. The variable (HIV05) was used as a sampling weight: further details can be found in the survey report [15]. The design was further stratified by states. The variable (V024) was used as the stratification variable. It is also important to specify the finite population correction (fpc) variable at each level. Not doing so implies that the sampling at the first level was done with replacement, thus ignoring all further levels. As the study uses data from a national survey, the sampling technique used would be without replacement. Variables (fpc_psu) & (fpc_hh) was used to calculate the fpc for PSU and household respectively. The FPC for the household is $\sqrt{N - n/N - 1}$ where N is the total number of households and n is the number of households in each PSU. Similarly, the FPC for PSU is calculated by setting (N) as the total number of PSUs and (n) as the number of PSUs in each state.

A two-stage approach for analysis was employed. In the first stage, the variables to be retained in the final model were identified. In the second stage, a 3 level multilevel model was fit using these variables and inference was drawn on this model. Similar approaches have been used in a variety of population health research settings [19,20]. Alternative model selection procedures exist and have been applied in the population health context, for example, the LASSO or penalized least squares approach [21–23]. These could be investigated in future work on this data.

Table 1. Distribution of variables in the study according to HIV status.

Variable	*HIV negative (N = 105 177)* Mean (SE) or N (%)	*HIV positive (N = 480)* Mean (SE) or N (%)	*p-value*
Age*	31.25 (0.01)	30.9 (0.12)	0.006
No. of living children*	1.94 (0.003)	2.0 (0.02)	0.003
Gender			
Female	39,043 (37.1)	214 (44.6)	0.001
Male	66,134 (62.9)	266 (55.4)	
Knowledge of contraceptives			
Knows no method	1693 (1.6)	14 (2.9)	0.06
Knows folkloric/traditional method	108 (0.1)	0	
Knows modern method	103,376 (98.3)	466 (97.1)	
Ever used contraceptive methods			
Never used	48,933 (46.5)	232 (48.3)	0.27
Used only folkloric/traditional method	7230 (6.9)	23 (4.8)	
Used modern methods	49,014 (46.6)	225 (46.9)	
Number of other partners spouse has			
No other partner	38,291 (48.6)	207 (55.5)	0.024
1 other partner	40,043 (50.8)	165 (44.2)	
$>$ 1 other partner	472 (0.6)	1 (0.3)	
Age at first marriage*	20.43 (0.12)	19.53 (0.05)	$<$ 0.001
Reduce risk of getting AIDS by not having sex at all			
No	9,917 (11.5)	51 (13.4)	0.35
Yes	66,349 (76.6)	293 (77.1)	
Don't Know	10,328 (12.0)	36 (9.5)	
Reduce risk of getting AIDS by using condoms			
No	7,551 (8.7)	25 (6.6)	0.27
Yes	66, 546 (76.9)	291 (76.6)	
Don't Know	12,497 (14.4)	64 (16.8)	
Reduce risk of getting AIDS by having only 1 sexual partner			
No	5, 236 (6.1)	27 (7.1)	0.78
Yes	72,104 (83.3)	316 (83.2)	
Don't Know	9,244 (10.7)	37 (9.7)	
Place of Residence			
Urban	49,627 (47.2)	234 (48.75)	0.49
Rural	55,550 (52.8)	246 (51.25)	
Marital Status			
Never Married	25,404 (24.2)	102 (21.25)	0.33
Ever Married	79,773 (75.9)	378 (78.75)	
Had an STI in last 12 months			
No	104,525 (99.4)	474 (98.8)	0.26
Yes	652 (0.6)	6 (1.3)	
PSU covered by Anganwadi			
Yes	30,664 (29.2)	152 (31.7)	0.47
No	74,496 (70.8)	328 (68.3)	
Wealth Index			
Poorest	10,201 (9.7)	37 (7.7)	0.43
Poorer	14,795 (14.1)	65 (13.5)	
Middle	21,339 (20.3)	92 (19.2)	
Richer	27,167 (25.8)	127 (26.5)	
Richest	31,675 (30.1)	159 (33.1)	

Notations: *Continuous variable, NC: variable information was not collected for the study.
Abbreviations used: SE: standard error, STI: Sexually Transmitted Infections AIDS: Acquired Immune Deficiency Syndrome.

All variables were summarized after accounting for the multi-level design and sampling weights. The distribution of all variables was compared across gender and HIV status (refer to Table 1). This was done to account for the fact that certain variables are only available for specific sexes. A bivariate method of analysis was used to identify the variables to retain in the final model. The odds ratios (ORs) obtained are the unadjusted ORs. Stata (version 14) was used for the same. Variables which showed significance ($p < 0.2$) in the bivariate

analysis were retained in the final model. This was done to allow more variables to be retained in the final analysis model by setting a less conservative p-value < 0.05. As the variable information collected from male and females vary, the analysis was stratified by gender. We also observe disparities in the prevalence of HIV between males and females. The overall prevalence of HIV was 0.557 for males and 0.442 for females. Furthermore, slightly different risk factors were measured on males and females so the analysis was done separately by gender to accommodate the differential prevalence and representation. We recognize that risk factors may likely to vary by residence status (urban/ rural); however, that is beyond the scope of this paper. Adjusted odds ratios (OR) of HIV status and their 95% confidence intervals (CI) were estimated after controlling for identified covariates. Further, the predictive ability of the model was checked using the area under the receiver operating characteristic curve (AUROC).

Finally, we used the intracluster correlation coefficient (ICC, denoted by ρ) [24] to assess the relatedness of each pair of variables within the clusters (at both PSU and household level). ICC is calculated as $\rho = \frac{\sigma^2{}_{(b)}}{\sigma^2{}_{(b)}+\sigma^2{}_{(w)}}$; where $\sigma^2{}_{(b)}$ is variability between clusters and $\sigma^2{}_{(\mathrm{w})}$ is the variability within clusters. A value of ICC close to 1 indicates high similarity between values within the same group, while a value close to zero implies low similarity within the group.

A logistic regression model for the survey data was then fit using the same variables as was included in the final model without accounting for the multiple levels in the dataset. This was done separately for both the sexes.

The standard errors (SE) for both models were compared. The effect of accounting for multiple levels over ignoring the different levels was noted. A similar strategy for variable selection was implemented when the model was fit after ignoring the multilevel structure. Variables that proved significant in the bivariate analysis were included in the final model. This was done to meet the aim of comparing the risk estimates obtained when the multilevel structure of the data is ignored. The simplest 3 level model i.e. the fully unconditional model is given by:

$$Y_{ijk} = \beta_{0jk} + e_{ijk} \tag{1.1}$$

$$\beta_{0jk} = \gamma_{00k} + r_{0jk} \tag{1.2}$$

$$\gamma_{00k} = \delta_{000} + u_{00k} \tag{1.3}$$

Combining the three equations,

$$Y_{ijk} = \delta_{000} + r_{0jk} + u_{00k} + e_{ijk} \tag{1.4}$$

where $(i = 1, 2 \ldots, n_{jk}), (j = 1, 2 \ldots, n_k)$ and $(k = 1, 2 \ldots, K)$; $e_{ijk} \sim N(0, \sigma^2)$, $r_{0jk} \sim N(0, \tau_r)$ and $u_{00k} \sim N(0, \tau_u)$.

The notations used here are adopted from Raudenbush and Bingenheimer [25]. A similar multilevel model was employed by Roux and Aiello [18] in their study. The model (1.2) is the household level model where we view each household mean (β_{0jk}) to be varying randomly around some district mean (γ_{00k}). Similarly, the model (1.3) is the district level model where each district mean (γ_{00k}) varies randomly around a grand mean (δ_{000}).

The model (1.4) represents how the total variation in the outcome variable is allocated across 3 levels i.e. Level 1: individuals within households,(σ^2); Level 2: households within districts (τ_r); and Level 3: among districts(τ_u); thus allowing the estimation of the proportion of variance at each level. This can be used to estimate the intra cluster correlation coefficient (ICC) denoted by (ρ). It is a measure of relatedness within clustered data.

The general level 1 model is given by:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{1ijk} + \beta_{2jk}X_{2ijk} + \ldots + \beta_{pjk}X_{pijk} + e_{ijk}$$

General Level 2 model:

$$\beta_{pjk} = \gamma_{p0k} + \sum_{q=1}^{Q} \gamma_{pqk}Z_{qjk} + r_{pjk}$$

Here, the variable Y_{ijk} can be interpreted as the variable measured for the i^{th} individual in the j^{th} household residing in the k$^{\text{th}}$ PSU. γ_{p0k} is the intercept for the k$^{\text{th}}$ district in modeling the household effect of β_{pjk}. Z_{qjk} is the vector of household level predictors for effect β_{pjk}.

General Level 3 model:

$$\gamma_{pqk} = \delta_{pq0} + \sum_{q=1}^{Q_p} \delta_{pqs}W_{sk} + u_{pqk}$$

where, δ_{pqs} is the corresponding district level coefficient that represents the direction and strength of association between W_{sk} and γ_{pqk}. W_{sk} is the level 3 characteristic used as a predictor for γ_{pqk}.

For logistic regression models, $Y_{ijk} = \log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right)$. Where π_{ijk} is the probability of the i^{th} individual in the j^{th} household of the k^{th} PSU being HIV positive.

4. Results

The proportion of male and female respondents in the study was 62.8% and 37.2% respectively. As previously noted, disparities in the prevalence of HIV between both sexes and a difference in the risk factors measured were observed for both males and females, and so the analysis was stratified by gender to accommodate this differential prevalence. Table 1 provides the distribution of all variables of interest by HIV status.

Table 2 summarizes the adjusted risk estimates associated with HIV status for males and females using multilevel mixed effect logistic regression approach (see Methods).

Among males, lack of knowledge of contraceptives, never used any contraceptive methods, STI in past 12 months, PSU was not covered by an Anganwadi, place of residence and the wealth index were all found associated with HIV risk. Married individuals were 1.3 (95% CI: 1.18, 1.37) times more at risk of HIV as compared to their counterparts. Further, males with a history of STI were 2.08 (95% CI: 1.67, 2.50) times more likely to be at risk of HIV as compared to those without. We also found a maximum OR of 2.37 (95% CI: 2.12, 2.67) for lack of knowledge of contraceptive methods. However, a protective association was obtained for residing in urban areas (OR = 0.79; 95% CI: 0.75, 0.84). This model also showed excellent predictive ability (Figure 1(a)).

Table 2. Odds Ratio (OR) and 95% CI for OR for factors influencing HIV status.

	Female ($N = 38{,}715$)		Male ($N = 66{,}212$)	
Variable	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
Age*	1 (0.99, 1)	0.90	0.99 (0.98, 1.00)	< 0.001
Number of living children*	1.01 (0.99, 1.03)	0.26	NS	NS
Knowledge of Contraceptive methods				
Knows modern methods ®	1		1	
Knows no method	1.35 (1.13, 1.60)	0.001	2.37 (2.12, 2.67)	< 0.001
Use of Contraceptive methods				
Used modern methods ®	1		1	
Never used	1.30 (1.22, 1.37)	< 0.001	0.91 (0.86, 0.96)	< 0.001
Hemoglobin levels (g/dl)*	1.0 (0.99, 1)	0.41	NS	NS
Number of other partners spouse has			NS	NS
No other partner ®	1			
$\geq$ 1 other partner	1.82 (1.49, 2.21)	< 0.001		
Age at first marriage*	1.03 (1.02, 2.21)	< 0.001	NS	NS
Had an STI in last 12 months				
No ®	1		1	
Yes	2.05 (1.67, 2.50)	< 0.001	1.49 (1.06, 2.09)	0.02
Had genital sore in last 12 months				
No ®	1		NS	NS
Yes	0.97 (0.79, 1.19)	0.80		
Had genital discharge in last 12 months			NS	NS
No ®	1			
Yes	1.59 (1.44, 1.74)	< 0.001		
Place of Residence	NS	NS		
Rural ®			1	
Urban			0.79 (0.75, 0.84)	< 0.001
Ever had blood transfusion			NS	NS
No ®	1			
Yes	0.486 (0.43, 0.55)	< 0.001		
Wealth Index				
Poorest ®	1		1	
Poorer	1.05 (0.93, 1.18)	0.434	1.66 (1.49, 1.84)	< 0.001
Middle	0.96 (0.85, 1.07)	0.459	1.55 (1.41, 1.70)	< 0.001
Richer	1.18 (1.04, 1.32)	0.007	1.52 (1.38, 1.66)	< 0.001
Richest	1.30 (1.15, 1.46)	< 0.001	1.54 (1.40, 1.70)	< 0.001
Marital Status	NS	NS		
Never Married ®			1	
Ever Married			1.3 (1.18, 1.37)	< 0.001
PSU covered by Anganwadi	NS	NS		
Yes ®			1	
No			1.57 (1.47, 1.67)	< 0.001
Random Effects	$\widehat{\sigma^2}$(95% CI)		$\widehat{\sigma^2}$(95% CI)	
PSU	2.910 (2.69, 3.14)		2.715 (2.53, 2.92)	
Household	7.461 (7.08, 7.85)		7.475 (7.16, 7.80)	

Notations: *Continuous variable, NC: variable information was not collected for the study, NS: Variable did not prove significant in bivariate analysis, R: Reference category.

Abbreviations used: OR: Adjusted Odds ratio; CI: confidence interval, g/dl: Grams per deciliter, STI: Sexually Transmitted Infections, and AIDS: Acquired Immune Deficiency Syndrome.

Among females, lack of knowledge of contraceptive methods, never used contraceptive methods, the spouse has one other partner, age at first marriage, having an STI or genital discharge in the last 12 months and wealth index categories (except the middle wealth index group) were observed as risk factors associated with HIV. We found the highest OR (2.05; 95% CI: 1.67–2.5) for having an STI in the last 12 months. Contrary to this, ever having had

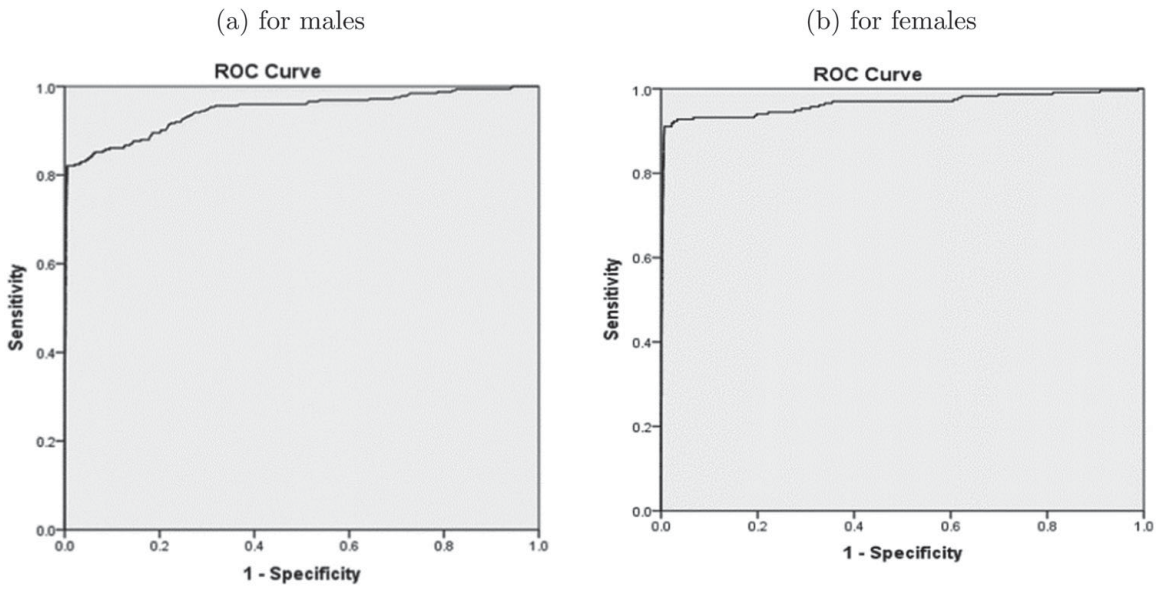


Figure 1. Receiver-Operator Characteristic (ROC) curves for final model.

a blood transfusion revealed a protective effect (OR = 0.486; 95% CI: 0.43, 0.55). Further, the AUROC from predictive probabilities showed an excellent predictive ability for this model (Figure 1(b)).

Adjusted risk estimates of HIV ignoring the hierarchical structure of data are provided in Table 3. We will now highlight the differences between the model ignoring the data structure and the model in Table 2 which accounts for the multilevel structure.

Among females, the risk estimates for the variables- knowledge on contraceptive methods, age at first marriage, and genital sore/genital discharge in the last 12 months were underestimated. The protective effect of number of living children (OR: 0.95 (95% CI: 0.94, 0.97)) and genital sore in the last 12 months (OR: 0.74 (95% CI: 0.63, 0.89)) are modified (OR: 1.01 (95% CI: 0.99, 1.03) and OR: 0.97 (95% CI: 0.79, 1.19)) respectively after accounting for multiple levels. A similar pattern was observed for knowledge of contraceptive methods among males. Further, variable indicating a lack of knowledge on contraceptive methods proved insignificant when ignoring multiple levels. Moreover, the risk estimates for variables pertaining to haemoglobin level, STI in last 12 months, residence status, wealth index, and marital status were overestimated due to ignoring the hierarchical nature of the data (Table 3).

The effect of using cluster sampling in studies can be quantified in another way through consideration of the intracluster correlation coefficient (ICC). We estimated the ICC values at PSU and household levels using the random effect modeling approach. We allowed for the possibility that the gender distribution of data may influence the ICC estimates and may present the values separately for males and females. Values of ICC are reported below and in Table 4; the associated 95% confidence intervals are included in Table 4.

In particular, for females, a high ICC between knowledge of contraceptives and HIV status was observed at both PSU and Household level (ICC_{psu}: 0.73, ICC_{hh}: 0.71). Similarly, a high correlation with HIV status is seen among variables pertaining to wealth index (ICC_{psu}: 0.63, ICC_{hh}: 0.72) and blood transfusion (ICC_{psu}: 0.68, ICC_{hh}: 0.85). Very

Table 3. Risk estimates of factors influencing HIV status ignoring the multiple levels.

	Females ($N = 38{,}715$)		Males ($N = 66{,}212$)	
Variable	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
Age*	1.006 (1.00, 1.01)	0.002	0.983 (0.98, 0.99)	< 0.001
Number of living children*	0.95 (0.94, 0.97)	< 0.001	NS	NS
Knowledge of Contraceptive methods				
Knows modern methods ®	1		1	
Knows no method	1.08 (0.91, 1.29)	0.391	0.98 (0.85, 1.13)	0.827
Use of Contraceptive methods				
Used modern methods ®	1		1	
Never used	1.48 (1.40, 1.56)	< 0.001	1.37 (1.25, 1.50)	< 0.001
Hemoglobin levels (g/dl)*	0.996 (0.994, 0.997)	< 0.001	NS	NS
Number of other partners spouse has			NS	NS
No other partner ®	1			
≥ 1 other partner	2.02 (1.77, 2.30)	< 0.001		
Age at first marriage*	0.988 (0.98, 0.99)	0.003	NS	NS
Had an STI in last 12 months				
No ®	1		1	
Yes	1.27 (1.00, 1.61)	0.047	1.59 (1.19, 2.13)	0.002
Had genital sore in last 12 months			NS	NS
No ®	1			
Yes	0.75 (0.63, 0.89)	0.001		
Had genital discharge in last 12 months			NS	NS
No ®	1			
Yes	1.27 (1.15, 1.40)	< 0.001		
Place of Residence				
Rural ®	1		1	
Urban	1.591 (1.49, 1.70)	< 0.001	1.23 (1.13, 1.35)	< 0.001
Ever had blood transfusion				
No ®	1		NS	NS
Yes	0.24 (0.06, 0.98)	0.047		
Wealth Index				
Poorest ®	1		1	
Poorer	1.20 (1.06, 1.36)	0.003	1.79 (1.49, 2.14)	< 0.001
Middle	1.13 (1.01, 1.27)	0.027	1.67 (1.39, 2.01)	< 0.001
Richer	1.30 (1.16, 1.45)	< 0.001	1.02 (0.87, 1.20)	0.811
Richest	1.49 (1.33, 1.66)	< 0.001	1.63 (1.37, 1.94)	< 0.001
Marital Status	NS	NS		
Never Married ®			1	
Ever Married			2.015 (1.72, 2.36)	< 0.001
PSU covered by Anganwadi	NS	NS		
Yes ®			1	
No			1.53 (1.02, 2.29)	0.040

Notations: *Continuous variable, ® Reference Category, NC: variable information was not collected for the study, NS: Variable did not prove significant in bivariate analysis.
Abbreviations used: OR, Adjusted Odds ratio; CI: Confidence interval, STI: Sexually Transmitted infections.

low correlation was observed between age and HIV status (ICC_{psu}: 0.037, ICC_{hh}: 0.032). Among males, a high correlation was reported between wealth index and HIV status (ICC_{psu}: 0.73, ICC_{hh}: 0.71). The study results show consistency in the value of the ICC across different levels (PSU and household) in the data. The ICC values themselves range from 0.004 to 0.848 (Table 4). However, even the smallest values of ICC can have a significant impact on estimates and standard errors as can be seen in the differences between Table 2 and Table 3.

Table 4. Intracluster correlation coefficients (ICC) for variables influencing HIV at PSU and Household levels.

	Female (N = 38,715)		Male (N = 66,212)	
Variable	PSU-ICC (95% CI)	Household-ICC (95% CI)	PSU-ICC (95% CI)	Household-ICC (95% CI)
Age*	0.037 (0.03, 0.04)	0.032 (0.02, 0.05)	0.0022 (0.000, 0.006)	⊙
Number of living children*	0.121 (0.11, 0.13)	0.119 (0.11, 0.13)	NS	NS
Knowledge of contraceptives	0.725 (0.63, 0.80)	0.705 (0.66, 0.74)	0.501 (0.46, 0.55)	0.532 (0.48, 0.58)
Use of Contraceptives	0.177 (0.16, 0.20)	0.172 (0.15, 0.19)	0.102 (0.09, 0.11)	0.084 (0.07, 0.10)
Wealth Index	0.631 (0.61, 0.65)	0.720 (0.70, 0.74)	0.532 (0.50, 0.56)	0.622 (0.60, 0.65)
Age at first marriage*	0.270 (0.26, 0.28)	0.268 (0.26, 0.28)	NS	NS
Had any STI in last 12 months	0.471 (0.41, 0.53)	0.411 (0.33, 0.49)	0.455 (0.35, 0.56)	0.427 (0.25, 0.62)
Had genital sore in last 12 months	0.307 (0.26, 0.36)	0.269 (0.20, 0.35)	NS	NS
Had genital discharge in last 12 months	0.299 (0.27, 0.35)	0.283 (0.25, 0.32)	NS	NS
Ever had blood transfusion	0.681 (0.65, 0.71)	0.848 (0.84, 0.85)	NS	NS
Marital Status	NS	NS	0.039 (0.03, 0.05)	0.022 (0.01, 0.04)

Notations: *Continuous variable, ⊙ : Truncated at zero.
Abbreviations used: PSU: Primary Sampling Units, NS: Variable did not prove significant in bivariate analysis, STI: Sexually Transmitted infections.

5. Discussion

Although there are many studies addressing the factors associated with HIV in the Indian population, the importance of accounting for the complex sampling strategy in the analysis has not been completely recognized. This study addresses this structure using a multilevel model for calculating the risk estimates for factors associated with HIV infection.

Among females, we found that individuals having no knowledge of the contraceptives and not using contraceptives during intercourse are 1.35 and 1.30 times more at risk of HIV respectively. Despite the protective effect of haemoglobin levels among females, some have reported that haemoglobin levels can be used as a predictive indicator for the progression of HIV/AIDS [5,26] and management of HIV [27].

Women with an STI within the last 12 months are 2.05 times more at risk of developing HIV compared to their counterparts. Vaginal infections are often characterized by genital discharge and are diagnosed among women with STIs. Bacterial vaginosis (BV) is the most prevalent cause of the vaginal discharge and is known to recur in higher frequency among HIV positive women. Those infected have shown to be at 1.59 times more risk of HIV compared to those who were not infected.

We observed a significant variation in the risk estimates after discarding the multilevel structure of the data. Variables pertaining to no knowledge of contraceptive methods, age at first marriage, wealth index and non-coverage of PSUs by anganwadis were significant risk factors for HIV when the multileveled model was used. A recent paper used the NFHS-3 data to report risk estimates for factors influencing HIV in India without accounting for the multiple levels in the data and the sampling methodology employed during data collection [4].

While the study helped identify a high-risk profile for HIV in India, it fails to account for the structure of the data. In our study, we obtained risk estimates for additional variables

pertaining to age at first marriage, the number of living children, history of blood transfusions and coverage of PSUs by anganwadis. Contrary to the earlier study [4], we observed that variables pertaining to the history of STI & genital discharge among females, use of contraceptives, marital status, and wealth index showed significant influence on HIV when a multilevel approach was employed. Finally, we identified that the perceived level of awareness of methods to reduce HIV, particularly among those who know not more than one method serves as a major threat to public health and demands serious attention.

A significant irregularity in the results of this study was obtained for women in the variable *ever had a blood transfusion*. Blood transfusion is proved a risk factor for transmission of HIV [3]; whereas the data reported ever having had a blood transfusion as a protective factor against HIV. Individuals are often screened for HIV before transfusion. Blood transfusions could be a good screening tool, thus reducing the odds of transmission of HIV via blood. However, recent reports suggest that transmission of HIV due to blood transfusions in India is on the rise [28,29].

6. Strengths and Limitations of this study

An inherent drawback of the cross-sectional data is its snapshot nature, which makes it challenging to establish a sequence to the occurrence of events and impossible to draw causal inference. Although NFHS is a large scale survey, it limits causal inference on the factors influencing HIV because of the cross-sectional nature of the design. Further, our analysis is restricted only to individuals who consented to HIV testing.

Even though NFHS is considered a high quality dataset, the logical outcome of investigating a large number of potential risk factors was the relatively high proportion of missing data, that ultimately resulted in discarding some of these risk factors. Variables pertaining to sterilization, age at first intercourse and time since last menstrual period were dropped from the analysis due to magnitude of missing data. Instead, HIV weights provided in the NFHS data was used to adjust for non-response rates in the survey. Nevertheless, the large-scale and nationally representative nature of the NFHS study along with its high response rates (women: 94.5%, men: 84.9%) and high coverage of health issues, make this an ideal source for a broad analysis of risk factors.

However, these concerns are counterbalanced by the study's strengths especially the large sample size and representativeness of the population which helps in generalizing the findings. Moreover, data were collected using uniform methods, standard tools and outcome measurements across the different states. Furthermore, we adopted a multilevel approach and accounted for the survey weights, a better choice for analyzing data of the hierarchical nature, to obtain reliable risk estimates of HIV infection in India.

7. Implications for research and practice

Several studies have highlighted the relevance of a multilevel approach than a single level approach to survey analysis [25,30]. We found a substantial difference in risk estimates among both males and females after accounting for the hierarchical nature of data. This further emphasizes the need to account for the structure of the data in order to obtain reliable risk estimates. In conclusion, this study has identified the risk profile for HIV infection using a modeling strategy appropriate to the data and has highlighted the changes in the

inference that arise when the structure is ignored. It offers a useful methodological guide towards both appropriate modeling strategies; an applied approach to the identification of future risk; and the direction needed to customize intervention to address HIV infection in the Indian population. In addition to improving awareness, focusing on behavior change strategies needs to be urgently addressed and implemented to reduce the burden of HIV in India.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Ms Nidhi Menon is a Ph.D. candidate at the Research School of Population Health at the Australian National University. Her research interests include multiple imputation, multilevel models and epidemiology.

Dr. Binukumar Bhaskarapillai is an Associate Professor, Department of Biostatistics, National Institute of Mental Health & Neuro Sciences (NIMHANS), Bengaluru, India. He has been involved in teaching Biostatistics postgraduates and research of interest area includes Multivariate Methods, Categorical data analysis, Missing data analysis, Statistical epidemiology and Meta analysis.

Dr. Alice Richardson is biostatistician and leader of the Data Analysis in Population Health Hub in the Research School of Population Health at the Australian National University. Her research interests encompass many aspects of biostatistics including multilevel modeling, robust statistics and multiple imputation. She also investigates innovative classroom techniques to discover the most engaging ways to convey statistical concepts to non-statisticians.

ORCID

Nidhi Menon http://orcid.org/0000-0001-8400-6079
Binukumar Bhaskarapillai http://orcid.org/0000-0003-3056-941X
Alice Richardson http://orcid.org/0000-0001-7084-1524

References

[1] International Institute for Population Sciences (IIPS) and Macro International. National family health survey (NFHS-3), 2005–06, India: volume I. Mumbai: IIPS; 2007.
[2] Tanwar S, Rewari B, Rao CD, et al. India's HIV programme: successes and challenges. J Virus Erad. 2016;2(Suppl 4):15–19.
[3] Godbole S, Mehendale S. HIV/AIDS epidemic in India: risk factors, risk behaviour & strategies for prevention & control. Indian J Med Res. 2005;121:356–368.
[4] Hazarika I. Risk factors for HIV-1 infection in India: evidence from the national family health survey. Int J STD AIDS. 2012;23(10):729–735.
[5] Obirikorang C, Yeboah FA. Blood haemoglobin measurement as a predictive indicator for the progression of HIV/AIDS in resource-limited setting. J Biomed Sci. 2009;16:102.
[6] Dletrich U, Maniar JK, Rübsamen-Waigmann H. The epidemiology of HIV in India. Trends Microbiol. 1995;3(1):17–21.
[7] Kumarasamy N, Vallabhaneni S, Flanigan TP, et al. Clinical profile of HIV in India. Indian J Med Res. 2005;121(4):377–394.
[8] Ramanathan S, Nagarajan K, Ramakrishnan L, et al. Inconsistent condom use by male clients during anal intercourse with occasional and regular female sex workers (FSWs): survey findings from southern states of India. BMJ Open. 2014 19;4(11):e005166.

[9] Bollinger RC, Brookmeyer RS, Mehendale SM, et al. Risk factors and clinical presentation of acute primary HIV infection in India. JAMA. 1997;278(23):2085–2089.
[10] Narain JP, Jha A, Lal S, et al. Risk factors for HIV transmission in India. AIDS. 1994;8(Suppl 2):S77–S82.
[11] Rodrigues JJ, Mehendale SM, Shepherd ME, et al. Risk factors for HIV infection in people attending clinics for sexually transmitted diseases in India. Br Med J. 1995;311(7000):283–286.
[12] Solomon S, Kumarasamy N, Ganesh AK, et al. Prevalence and risk factors of HIV-1 and HIV-2 infection in urban and rural areas in Tamil Nadu, India. Int J STD AIDS. 1998;9:98–103.
[13] Panda S, Kumar MS, Lokabiraman S, et al. Risk factors for HIV infection in injection drug users and evidence for onward transmission of HIV to their sexual partners in Chennai, India. J Acquir Immune Defic Syndr. 2005;39(1):9–15.
[14] Sogarwal R, Bachani D. Awareness of women about STDs, HIV/AIDS and condom use in India: lessons for preventive programmes. Heal Popul Perspect Issues. 2009;32(3):148–158.
[15] National Family Health Survey, India. [cited 2015 Apr 12]. Available from http://www.rchiips.org/nfhs/nfhs3.shtml.
[16] Pandey A, Reddy DCS, Ghys PD, et al. Improved estimates of India's HIV burden in 2006. Indian J Med Res. 2009;129(1):50–58.
[17] Ng M, Gakidou E, Levin-Rector A, et al. Assessment of population-level effect of Avahan, an HIV-prevention initiative in India. Lancet. 2011;378(9803):1643–1652.
[18] Roux AVD, Aiello AE. Multilevel analysis of infectious diseases. J Infect Dis. 2005;191(S1): S25–S33.
[19] Zahangir MS, Hasan MM, Richardson A, et al. Malnutrition and non-communicable diseases among Bangladeshi women: an urban–rural comparison. Nutr Diabetes. 2017;7:e250.
[20] Hasan MM, Richardson A. How sustainable household environment and knowledge of healthy practices relate to childhood morbidity in South Asia: analysis of survey data from Bangladesh, Nepal and Pakistan. BMJ Open. 2017;7:e015019.
[21] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
[22] Masud A, Tu W, Yu Z. Variable selection for mixture and promotion time cure rate models. Stat Methods Med Res. 2018;27:2185–2199.
[23] Li Z, Liu H, Tu W. A sexually transmitted infection screeig algorithm based on semiparametric regression models. Stat Med. 2015;34:2844–2857.
[24] Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Ann Fam Med. 2004;2(3):204–208.
[25] Bingenheimer JB, Raudenbush W. Statistical methods and substantive inferences in public health: issues in the application of multilevel models. Annu Rev Public Health. 2004;25:53–77.
[26] De Santis GC, Brunetta DM, Vilar FC, et al. Hematological abnormalities in HIV-infected patients. Int J Infect Dis. 2011;15(12):e808–e811.
[27] Shelton J D, Cassell M M, Adetunji J. Is poverty or wealth at the root of HIV? Lancet. 2005;366(9491):1057–1058.
[28] Roy SD. 1,000 HIV+ cases in Maharashtra due to infected blood transfusion. [cited 2015 May 10]. Available from http://timesofindia.indiatimes.com/india/1000-HIV-cases-in-Maharashtra-due-to-infected-blood-transfusion/articleshow/46047954.cms
[29] Rohit PS. HIV via blood transfusion on the rise. cited [2015 May 10]. Available from http://timesofindia.indiatimes.com/cityhyderabad/HIV-via-blood-transfusion-on-the-rise/articleshow/21926970.cms
[30] Guo G, Zhao H. Multilevel modeling for binary data. Annu Rev Sociol. 2000;26:441–462.

7.4 Case Study 3: Developing a rich imputation model for imputing laboratory data

7.4.1 Introduction

Chapter 2 discussed the need to focus on developing rich imputation models, and the methods to select predictors for the imputation model. Key takeaways from this chapter were that the imputation model should not only reflect the sampling strategy employed in the study, but also be developed on the basis of sophisticated quantitative measures. Uncongeniality between imputation and analysis models (see section 2.3.8 for more details) can occur if the imputer and the analyst are not the same person and quantitative methods are not used to develop the imputation model. This could be because different assumptions are made when constructing the imputation and analysis model. Thus there is a need to standardise and quantify the process to try and minimise subjectivity in developing the imputation model.

This case study aims to provide an analytical investigation of the laboratory diagnosis of Hepatitis C (HCV) infection and focus on how to maximize the predictive value of routine pathology data. In this study, we recommend using the Influx - Outflux measures to help construct the imputation model when using multiple imputation.
These missing data pattern statistics were proposed by (van Buuren, 2018). However, there is a gaping hole in the current literature on the implementation of these statistics in constructing an imputation model for real world data. In this study, we use the data that was made available by ACT Pathology at the Canberra Hospital, Canberra, Australia to illustrate the use of Influx and Outflux in missing data imputation.
The study is novel in its approach and is among the first to demonstrate the application of missing data pattern statistics in routine pathology variables to increase diagnosis of Hepatitis C infection.

The Effect of Multiple Imputation of Routine Pathology Variables on Laboratory Diagnosis of Hepatitis C Infection

Nidhi Menon*, Brett A. Lidbury & Alice Richardson

National Centre for Epidemiology & Population Health, Australian National University, Canberra, Australia

Abstract

Background: Pathology tests are central to modern healthcare in terms of diagnosis and patient management. Aggregated pathology results provide opportunities for research into fundamental and applied questions in health and medicine, but data analytic challenges appear since test profiles vary between medical practitioners, resulting in missing data. In this study we provide an analytical investigation of the laboratory diagnosis of Hepatitis C (HCV) infection and focus on how to maximize the predictive value of routine pathology data. We recommend implementing multiple imputations in combination with the Influx - Outflux values to help identify predictors in the presence of missing data.

Methods: Data from 14,320 community-patients aged 15 - 100 years were accessed via ACT Pathology (The Canberra Hospital, Australia). Influx and Outflux were calculated to identify which variables were potentially powerful predictors of missing values. Available Case analysis and Multiple Imputation were used to accommodate missing values in the dataset. Logistic regression model and stepwise selection method were used for analysing the imputed datasets. The predictive power

*Corresponding author at: National Centre for Epidemiology & Population Health, Australian National University, Canberra, Australia
email address: nidhi.menon@anu.edu.au

of all methods was compared.

Results: The predictive power of the models on multiply imputed data was similar to the power of the models based on complete data. The advantage of multiply imputed data was that it allowed for the inclusion of all the completed variables in the logistic models, thus identifying a broader selection of test results that could lead to the enhanced laboratory prediction of HCV.

Conclusions: Multiple imputation is an important statistical resource allowing all individuals in a study to contribute whatever data they have supplied to the analysis. Age, gender and alanine aminotransferase have been shown to be strong laboratory predictors of HCV infection.

Keywords: Hepatitis; logistic regression; multiple imputation.

Highlights:

- Multiple imputation of missing data is a well-established technique in statistics.
- Laboratory diagnosis of Hepatitis C infection can be achieved using routine pathology data.
- Areas under ROC curve achieved in our analysis were around 70%
- Multiple imputation in combination with the Influx - Outflux allows novel variables to contribute to the prediction models
- Age, gender and alanine aminotransferase are strong laboratory predictors of HCV infection.

1 Introduction

Pathology laboratories generate large quantities of data representing human function, including analyses of blood chemistry and cells, infections, genetics and more. These data represent an under-utilized research resource, and while there are drawbacks [1], they introduce a rich informatics and statistical substrate to assist clinical decision making. They also support the interrogation of databases to assist answering research problems; for example, such databases have been successfully mined for enhanced laboratory prediction of infectious diseases [2, 3]. Given the observational nature of the data collected, one drawback is that not every test is conducted for every patient. Therefore, pathology databases contain many missing values that potentially dilute the value of the data base.

Missingness can be one of three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are MCAR if the probability of missingness is not related to the value of the observation or any other variables in the data. For most datasets, the assumption of MCAR is unlikely to be satisfied unless the data is missing due to the design employed. If the missingness depends on the value of another variable in the dataset, then the data is said to be MAR. MCAR is a special case of MAR, thus if the data is MCAR, they are also MAR. If the MAR assumption is violated, then the data is said to be MNAR [4].

Missing data mechanisms are important since older methods of handling missing data (e.g. available case analysis) assume MCAR missingness while modern techniques to handle missing data such as maximum likelihood estimation (MLE) and multiple imputation assume only MAR.

1.1 Hepatitis C Virus

Hepatitis virus is a world-wide health concern, with an estimated 1.45 million deaths attributable to the virus globally in 2013 [5]. More recent modelling [6] has also shown that the global prevalence of Hepatitis C virus (HCV) has reached an estimated 71.1 million infections. HCV affects the liver and can cause permanent and ultimately fatal liver cancer [7]. HCV causes the inflammation of the liver and over time this can lead to liver diseases like cirrhosis and fibrosis. HCV is a blood borne virus and in developed coun-

tries it is often transmitted through risky behaviour such as sharing needles. Indeed, in Australia in 2007 [8], 90% of the new and 80% of the existing infections are transmitted in this way.

While there are advances in the treatment of people with HCV, the human costs of HCV, in terms of reduction of quality of life and well being and through occupational and social discrimination and isolation, remain significant [9]. Guidelines for the care and treatment of HCV have recently been updated [10, 11]. The financial costs of the virus for medical and hospital care, lost productivity, and the need for social support are also an increasing proportion of healthcare expenditure in Australia. While the diagnosis of HCV is most confidently made through the immunoassay [11], having a powerful predictive model will help focus on early detection of HCV infection particularly for cases where the specific immunoassay has not been ordered.

1.2 Multiple imputation

Analysis of multivariate data is often negatively affected by incompleteness. Reduced statistical power, increase in standard errors, complications in data handling and analysis, and introduction of bias are some of the common problems associated with missing data [12].

Several methods have been proposed over the years to handle the problem of missing data, for example single imputation using means or medians [13] and the EM algorithm [14]. While single imputation techniques are easy to implement, they treat the missing values as if they were known, thus eliminating any uncertainty that missing values bring to the dataset. They do not preserve the inherent variability of the imputed data and can be severely biased [14]. For these reasons we will not be pursuing Single Imputation in this paper but instead show the usefulness of Multiple Imputation (MI) [15], where the multiple imputed datasets are analysed using standard statistical procedures and the estimates from these analyses are then combined to produce overall estimates with more appropriate standard errors. These estimates and standard errors reflect the true variation and uncertainty in the data better than the estimates obtained by deleting any observations with missing values, also known as available case analysis. MI is less biased than single imputation methods, increases the efficiency of estimation and also preserves the variability in the dataset. Hence, it is the preferred method

to fill in missing values in datasets such as the one used in this study. It should be noted that the goal of imputation is not to replace or recover these missing values from the dataset, as they are unknown, but to produce valid and analytical results in the presence of the missing values.

Despite its conception in the 1980s, MI has only recently come into general use across medical research due to advances in statistical computation. Even with the evolution of MI and its statistical advantages, the technique has had limited application in laboratory diagnostics. The use of MI in clinical chemistry has been supported [16, 17] when values of a variable in an existing prediction model are missing. The study presented herein extends this previous approach by imputing variables with an unknown relationship to the outcome of interest.

The primary goal of this study is to show the importance of multiple imputation when dealing with missing values in routinely collected pathology data. A secondary goal of this paper is to contribute to knowledge of laboratory prediction of HCV by identifying novel combinations of routinely measured biomarkers that are highly predictive of HCV. This focus on an existing resource which provides a low-cost approach to knowledge discovery is important in the context of low-middle income countries, where laboratory resources may be limited [18].

2 Materials and Methods

2.1 Data

The dataset employed in this study was made available by ACT Pathology at The Canberra Hospital, Canberra, Australia. Patient identifiers were removed and only laboratory ID numbers provided. The data set contained the 18,625 pathology requests between 1997 and 2007 that included a request for either the assay for Hepatitis B (HepB) or Hepatitis C (HepC) or both (Figure 1). There were 4,296 patients for whom 60% or more of the other assays (Table 1) were missing, and these were removed before analysis commenced due to the large proportion of missingness. Another 3,546 individuals were missing a HCV immunoassay (HepC) outcome despite a request having been made and so excluded from the analysis. The result of the HCV

immunoassay is recorded as positive if the test is positive to the presence of HCV anti-bodies. The same dataset has been analysed previously [2, 3] to ascertain the interaction between virus, outcome, pre-processing and method on the performance of decision tree ensembles and support vector machines.

For access to de-identified patient data, this study had human ethics approval granted by The Australian National University Human Ethics Committee (2012/349) and the ACT Health Human Research Ethics Committee (ETHLR.11.016).

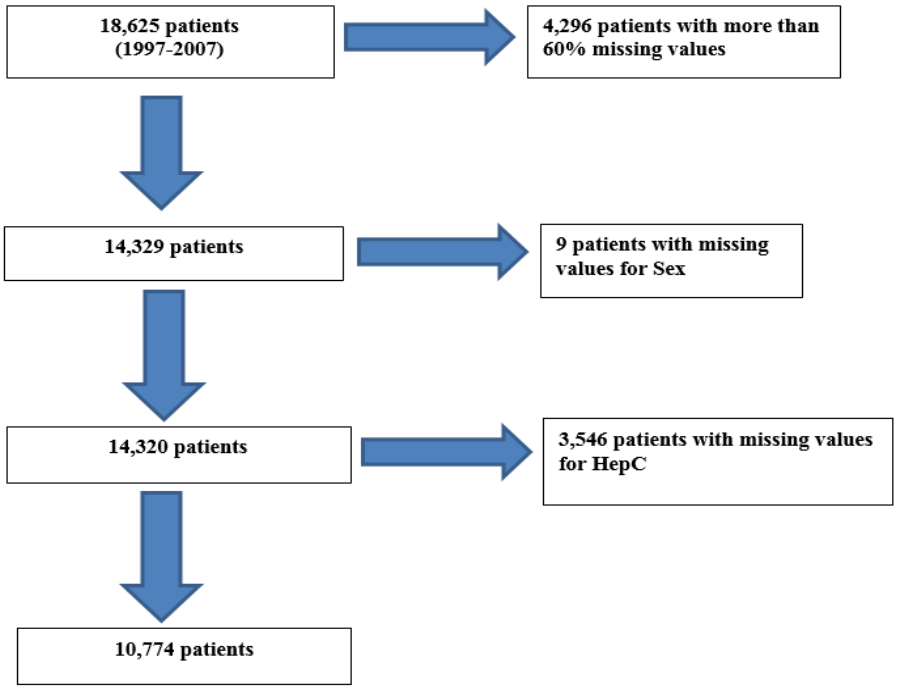


Figure 1: Flow of patients into the HepC study *(To be printed in Colour)*

Table 1: Diagnostic pathology variables included in the regression analysis, with summary statistics (n = 10,774).

Variable	*Description*	*Mean (SD) (HepC = 0) (n = 10102)*	*Mean (SD) (HepC = 1) (n = 672)*	*p-value*
Age	Patient's age in years	44.72 (19.36)	40.09 (14.41)	< 0.001
Sex	Patient's gender (Males) Frequency (%)	5188 (51.36%)	251 (37.35%)	0.00013
ALB	Albumin	42.84 (5.80)	42.09 (5.88)	0.0004
Sodium	Sodium	139.63 (3.22)	139.66 (3.06)	0.9927
K	Potassium	4.00 (0.45)	4.07 (0.47)	< 0.001
RCC	Red cell count	4.53 (0.64)	4.68 (0.64)	< 0.001
RDW	Red cell distribution width	13.88 (1.78)	14.07 (1.63)	< 0.001
ALT	Alanine aminotransferase	1.46 (0.38)	1.64 (0.42)	< 0.001
ALKP	Alkaline Phosphate	1.92 (0.21)	1.94 (0.18)	< 0.001
Crea	Creatinine	1.93 (0.16)	1.90 (0.13)	< 0.001
GGT	Gamma-glutamyl transferase	1.61 (0.44)	1.69 (0.43)	< 0.001
Urea	Blood urea	0.72 (0.20)	0.65 (0.19)	< 0.001
Plt	Platelets	2.39 (0.20)	2.39 (0.18)	0.2179
WCC	White cell count	0.87 (0.17)	0.90 (0.15)	< 0.001
TBil	Total bilirubin	1.03 (0.30)	0.98 (0.31)	< 0.001
Mono	Monocytes	0.74 (0.20)	0.76 (0.15)	< 0.001
Eos	Eosinophils	0.38 (0.18)	0.39 (0.18)	0.014
Bas	Basophil	0.18 (0.09)	0.19 (0.08)	< 0.001
Lymph	Lymphocytes	1.38 (0.34)	1.48 (0.53)	< 0.001

2.2 Statistical Analysis

This is a cross-sectional dataset where each individual only appears once with a non-monotone pattern of missingness. Percentage missingness is calculated for each variable and we examine the pattern of missingness visually to get an impression of the extent of incompleteness. These patterns of missingness influences the amount of information that can be transferred between variables [19]. For example, imputations can be more precise if complete information is available for the other variables for the observations that are to be imputed.

These patterns can be used to define inbound and outbound statistics, where the inbound statistics measures how well the missing values in one variable (X_i) are connected to other (X_j); and the outbound statistic measures how well the observed values in X_i are connected to X_j. Influx and Outflux are two overall measures of variable connectivity [19]
The influx coefficient is defined as

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ij}) r_{ij}}{\sum_k^p \sum_i^n r_{ik}}$$

The outflux coefficient is given by

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n r_{ij} (1 - r_{ik})}{\sum_k^p \sum_i^n (1 - r_{ij})}$$

The Influx of a completely observed variable is equal to 0 while the influx of a completely missing variable is equal to 1. For two variables with the same amount of missing data, the variable with higher influx is better connected to the observed values and thus, may be easier to impute. Outflux of a completely observed variable is 1, and outflux of a completely missing variable is 0. It is an indicator of the possible effectiveness of variables for imputation of missing data. Variables with higher values of outflux are better associated to the missing values and can be used for imputing other variables [19]. Influx and Outflux was computed for all variables in the dataset, as well as the fraction of incomplete cases among cases with X_j observed (an outflux statistic denoted by FICO). A fluxplot (Figure 3) is used to aid in constructing a predictive model.

Multiple imputation is used to address missing data. In this method, each missing value is replaced by two or more imputed values to represent the statistical uncertainty about which value to impute. It involves three stages; Imputation, Analysis and Pooling. Final inference on the regression coefficient estimates is made on the pooled result using Rubin's combination rule [13]. Variables with an Outflux coefficient above 0.95 i.e. variables that occurred in the top left corner of the flux plot, are used as predictors in the imputation model. We employ a stepwise procedure by calculating the number of times a variable occurred a predictor in the imputation model. This method is also referred to as the impute then select method [20]. A supermodel is constructed using the variables that occurred across all imputed datasets. For all other variables, we use the multivariate Wald test and the likelihood ratio test to determine whether the variable should be included in the final model.

Multiple Imputation using Chained Equations (MICE) [19] has emerged in the literature as a routine method for handling missing data in both continuous and binary variables. Descriptive statistics (mean and standard deviation) for all iterations are used to obtain a plot to check imputation convergence. If the mean and standard deviation of the imputed variables appear to have settled at particular values, the imputation process is deemed to have converged. There is a lack of formal tests of convergence [21], so the plots of statistics such as the mean and standard deviation are used to provide visual information about convergence. A backwards step wise model selection procedure is employed in the analysis phase after imputations.

All analysis was performed using the statistical software R (version 3.4.0) [22]. The package 'mice' [23] were used for multiple imputation. The package 'ROCR' [24] is used to produce Receiver Operating Characteristic (ROC) curves to test the predictive ability of the imputed models and to calculate the area under the curve (AUC). The packages 'nortest' [25] and 'psych' [26] are used to perform the Anderson-Darling test of normality and to obtain summary statistics respectively. The package 'MASS' [27] is used to perform χ^2 and Mann-Whitney U tests.

Table 2: Diagnostic pathology variables included in the regression analysis, with summary statistics (n = 10,774).

Variable	*Percentage Missing (HepC = 0) (n = 10102)*	*Percentage Missing (HepC = 1) (n = 672)*
Age	0	0
Sex	0	0
Urea	18.71	13.39
TBil	18.71	4.46
ALT	18.67	4.46
Crea	18.65	13.39
GGT	18.63	4.02
Potassium	17.52	12.35
ALKP	17.43	3.57
Sodium	17.19	11.61
ALB	16.41	3.57
Bas	2.52	2.23
Eos	1.97	1.04
Plt	0.86	0.15
Mono	0.53	0.30
Neut	0.53	0.30
Lymph	0.53	0.30
WCC	0.50	0
MCHC	0.23	0
Hct	0.21	0
RCC	0.21	0
RDW	0.17	0.45
MCV	0.16	0
Mch	0.16	0
Hb	0.12	0

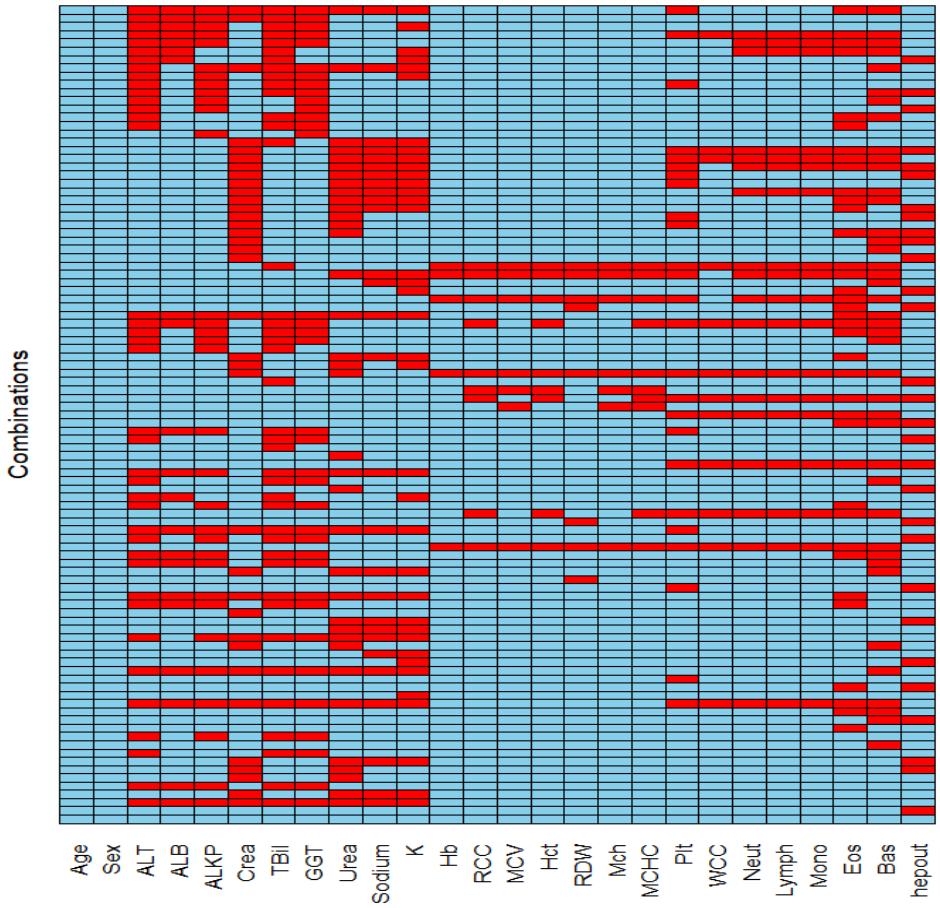


Figure 2: Patterns of missingness in predictor variables by combination. Each row of the chart represents a different combination of missing values.
Blue = a variable has no missing values, red = a variable has missing values.
See Table 2 for abbreviation of biomarker names.*(To be printed in Colour)*

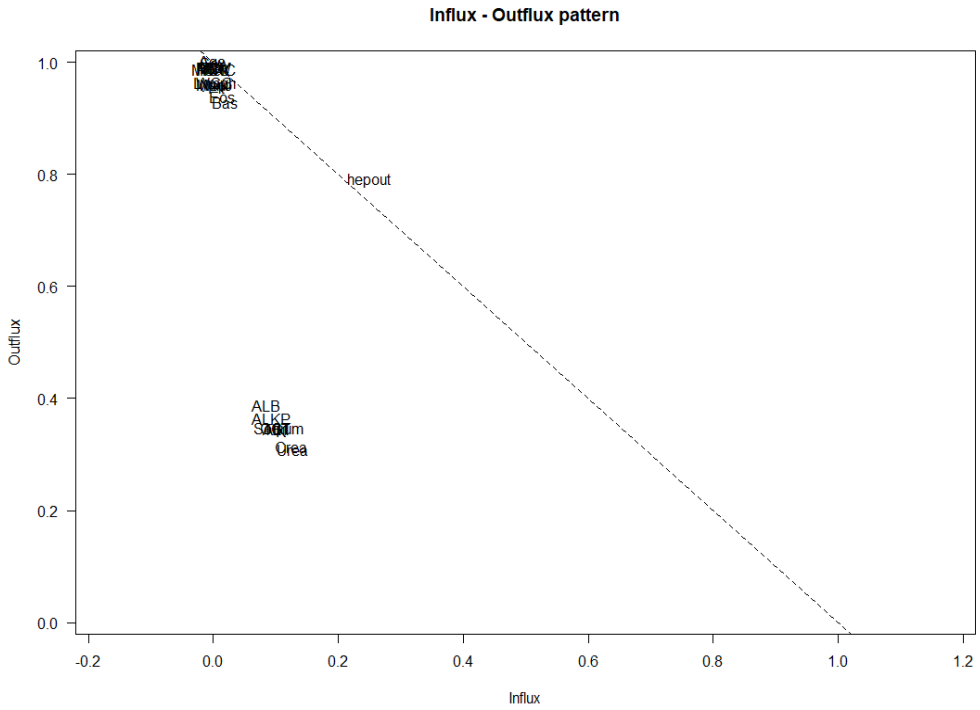


Figure 3: Fluxplot for HepC data: Outflux v/s Influx

Table 3: Influx and outflux of multivariate missing data patterns.

Variable	*Proportion Observed*	*Influx*	*Outflux*	*FICO*
Age	1.000	0.000	1.000	0.454
Sex	1.000	0.000	1.000	0.454
ALT	0.865	0.101	0.346	0.369
ALB	0.882	0.086	0.387	0.381
ALKP	0.875	0.093	0.365	0.376
Crea	0.840	0.126	0.314	0.350
TBil	0.865	0.102	0.345	0.369
GGT	0.866	0.101	0.348	0.369
Urea	0.839	0.127	0.310	0.349
Sodium	0.862	0.105	0.348	0.366
K	0.858	0.109	0.343	0.364
Hb	0.999	0.000	0.992	0.453
RCC	0.998	0.001	0.987	0.453
MCV	0.999	0.001	0.992	0.453
Hct	0.998	0.001	0.987	0.453
RDW	0.998	0.001	0.991	0.453
Mch	0.999	0.001	0.992	0.453
MCHC	0.998	0.001	0.987	0.453
Plt	0.993	0.005	0.956	0.450
WCC	0.996	0.002	0.964	0.452
Neut	0.996	0.002	0.961	0.451
Lymph	0.996	0.002	0.961	0.451
Mono	0.996	0.002	0.961	0.451
Eos	0.983	0.014	0.939	0.444
Bas	0.978	0.019	0.928	0.442
hepout	0.752	0.250	0.791	0.274

The variables are imputed on the raw scale. Log or square root transformations are applied to those variables with highly skewed distributions using passive imputation [28]. Passive imputation is a method to handle derived variables in imputation by transforming the imputed values of the variable. Even though a Normal distribution is not essential in the pre-

dictors, transformations were applied to ensure predictors were of similar order of magnitude across analysis methods. Complex transformations (like Box-Cox power transformations) were not considered for the final model for clarity of interpretation. Finally, logistic regression is used to predict HCV status on the basis of age, sex and the 23 biomarkers.

3 Results

There were over 100 different patterns of missingness in the predictor variables across the 14,320 patients (Figure 2a). There were 7,820 individuals (54.6%) with complete data on age, sex and all 23 biomarkers.
Percentages of missingness (Figure 2b, Table 2) occurs in one of three ways. For Sodium, Potassium, creatinine (Crea) and Urea, there is approximately 15% missingness, whether or not a patient was HCV positive. For serum albumin (ALB), alanine aminotransferase (ALT), alkaline phosphatase (ALP), gamma glutamyl transferase (GGT) and total serum bilirubin (TBil), there is about 15% missingness amongst HepC negative, and less than 1% missingness amongst HepC positive. For the remainder (Age, Sex, platelets (Plt), monocytes (Mono), eosinophils (Eos), basophils (Bas), lymphocytes (Lymph)), there was only occasional missingness of data.

The area under the ROC curve (AUROC) (Figure 5a) for the available cases logistic regression model on the validation dataset was 72%; namely, the model has a 72% chance of correctly classifying individuals as having a positive HepC assay. The model identified the variables Age, Sex, Sodium, RDW, Potassium, log(ALT), log(ALP), log(Crea), log(TBil), log(Urea) and sqrt(Lymph) as significant ($p < 0.05$; see Table 3).

As stated earlier, variables with higher outflux are potentially powerful predictors and can be used to impute variables with missing values. As indicated by the flux plot (Figure 3), variables in the far left corner have more complete data. Variables closer to the diagonal have more balanced values of influx and outflux. The group below the diagonal with higher values of influx depend highly on the imputation model. In this study, we consider variables with outflux values > 0.90 to be included in the imputation model. These

comprise: Age, Sex, Hb, RCC, MCV, Hct, RDW, Mch, MCHC, Plt ,WCC, Mono, Eos, Bas, Neut and Lymph.

We fit a stepwise logistic model to predict HepC and count the number of times each variable occurred in the imputation model in each of the imputed datasets. This was done for five, and 20 imputations. We observe that variables that did not occur in any of the imputation models when there were five imputations, occur in lower frequencies when there were 20 imputations.

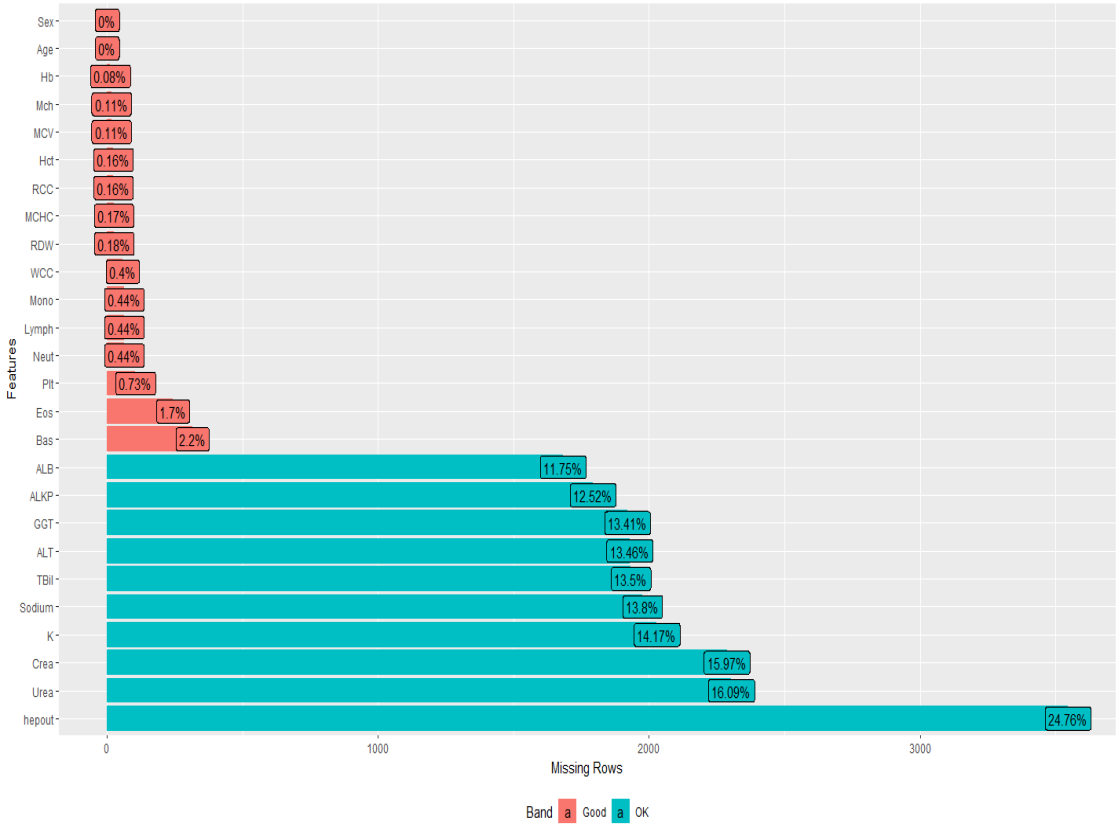


Figure 4: Percentage of missingness across variables in the dataset. Variables coloured pink have between 0 and 10% mising values, variables in blue have between 11 and 25% missing values. *(To be printed in Colour)*

Table 4: Tabulation of the number of occurrences in imputation model across all imputed datasets.

Variable	*No. of Occurrences (5 imputations)*	*No. of Occurrences (20 imputations)*
Age	5	20
ALB	5	20
ALT	5	20
Lymphocytes	5	20
RCC	5	20
RDW	5	20
Sex	5	20
TBil	5	20
Urea	4	20
Crea	5	18
K	3	15
MCV	4	14
Basophils	5	13
Hb	3	13
MCHC	3	13
Sodium	4	13
Hct	2	9
ALKP	0	7
Monocytes	2	6
Eosinophils	0	5
Mch	3	5
WCC	0	3
Neut	0	2
Plt	0	1

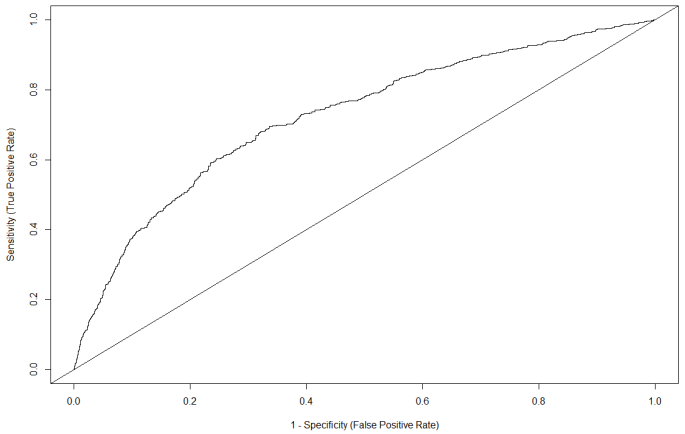
Variables that occurred in all imputation models were considered in the supermodel.For other variables, the likelihood ratio test was applied to identify if the variable should be included in the final model. Variables with p-value < 0.05 were selected in the final supermodel. Table 4 shows the results of the tabulation for number of occurrences of each variable when number of imputations were equal to 5 and 20.

Table 5: Regression coefficients for logistic regression models. Log = log transform applied before regression analysis. Sqrt = square root transformation applied before regression analysis

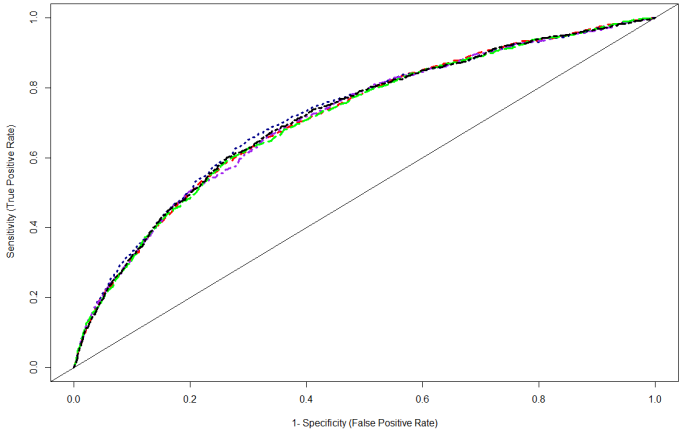
Variable	CC: $\beta(SE)$	CC: p-value	MICE (m = 5): $\beta(SE)$	MICE (m = 5): p-value	MICE (m = 20): $\beta(SE)$	MICE (m = 20): p-value
Intercept	3.408 (2.455)	0.1651	31.577 (8.937)	< 0.001	-0.7547 (3.2853)	0.8183
Age	-0.016 (0.003)	< 0.001	-0.016 (0.003)	< 0.001	-0.0169 (0.003)	< 0.001
Sex	0.664 (0.109)	< 0.001	0.516 (0.112)	< 0.001	0.5043 (0.0956)	< 0.001
RDW	0.087 (0.026)	< 0.001	0.142 (0.023)	< 0.001	0.1592 (0.0254)	< 0.001
Log(ALT)	1.415 (0.123)	< 0.001	0.3504 (0.0481)	< 0.001	0.3436 (0.0445)	< 0.001
Sqrt(Lymph)	0.312 (0.112)	0.0056	0.3912 (0.099)	< 0.001	0.3406 (0.0943)	< 0.001
ALB	-0.016 (0.009)	0.0644	-0.044 (0.008)	< 0.001	-0.0473 (0.0084)	< 0.001
Potassium	0.383 (0.112)	< 0.001	–	–	0.1439 (0.0962)	0.1346
Sodium	-0.035 (0.016)	0.0292	–	–	–	–
Sqrt(Bas)	0.701 (0.361)	0.0523	–	–	–	–
Log(ALKP)	-0.921 (0.269)	< 0.001	–	–	–	–
Log(Crea)	-1.320 (0.558)	0.0180	-0.3904 (0.195)	0.0046	–	–
Log(TBil)	-1.114 (0.170)	< 0.001	-0.3945 (0.065)	< 0.001	-0.4031 (0.0656)	< 0.001
Log(Urea)	-1.261 (0.338)	< 0.001	-0.251 (0.135)	0.0652	-0.4706 (0.114)	< 0.001
Sqrt(Mono)	–	–	-0.174(0.275)	0.5251	–	–
Hb	–	–	0.0913(0.022)	< 0.001	0.0774(0.0335)	0.0211
MCHC	–	–	-0.0902(0.024)	< 0.001	–	–
MCV	–	–	-0.2328 (0.0889)	0.0089	0.0712 (0.0549)	0.0194
RCC	–	–	-2.107 (0.698)	0.0025	-2.0800 (0.7416)	0.0051
Mch	–	–	0.503 (0.2762)	0.0688	-0.4026 (0.1620)	0.0131
Hct	–	–	–	–	4.3233(11.205)	0.0699

Figure 5: Receiver-Operator Characteristic (ROC) curves for each method of analysis. Diagonal line represents ROC of 50% *(To be printed in Colour)*

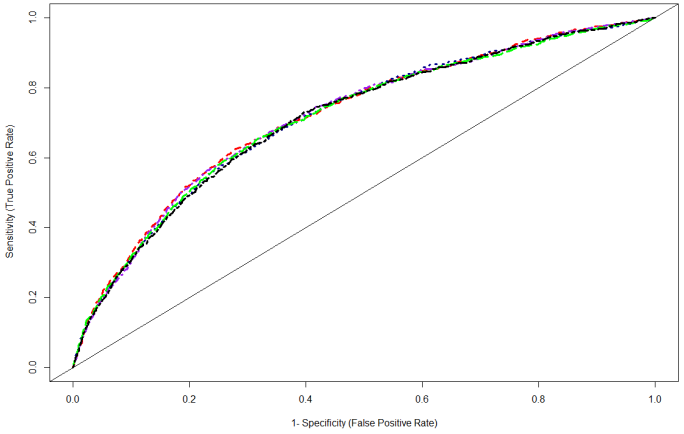
(a) Available Case analysis



(b) MICE (m = 5)



(c) MICE (m = 20)



The parameter estimate from the multiple imputation methods are automatically pooled and presented in Table 5. The AUROCs of the five-fold imputed models have a mean of 72% (SD 0.003). We increased the number of imputations from 5 to 20 and observed variable selection was more consistent, reducing the randomness in the variable selection process. The AUROCs of the twenty-fold imputed models have a mean of 71% (SD 0.002).

4 Discussion

Discussion of the results will involve both the meaning of the results for multiple imputation in general, and the results for the diagnosis of hepatitis C in particular. The focus of the study is to improve the diagnostic process by developing a powerful predictive model for HepC without discarding any information collected during laboratory tests due to the presence of missing values, and using the completed dataset for developing this model.

The percentage of missing values displayed three patterns (see Results). Biomarkers with around 15% missingness included sodium, potassium, creatinine and urea which are associated with the Urea-Electrolytes-Creatinine battery of tests. These are routinely requested for most laboratory investigations. Biomarkers such as ALB, ALT, ALP, GGT and TBil and displayed around 15% missingness for those found HCV negative, and less than 1% missingness for those found HCV positive. These are routine liver function tests and would be primary markers sought when HepC is suspected. The remaining biomarkers are from the full blood count and are routinely requested, and so records are complete for almost all individuals. A recent paper [29] has been published on this strategy, but with the goal of avoiding painful and expensive liver biopsy.

The multiply imputed analyses are largely consistent with the available case analysis [2, 3] (Table 3). However, three variables in the single imputation model were not identified as important disease markers in the analysis of the imputed datasets; hence some unnecessary variables were removed via MI. Although the ROC curves (Figure 5) produced by the MI methods did not yield high absolute accuracy, the ten key predictors common to all the multiply imputed regression models suggest a pattern of routine, easily accessible pathology markers, which predict a HCV infection. In addition to

early detection, the advantage of using quality routine pathology data as a predictive model, where immunoassay is not easily available is a key outcome especially when handling population health issues for rural or remote areas in third world countries.

The regression models suggest that increased levels of ALT, RDW and lymphocytes increased the odds of detecting HCV infection. All models presented include the variable ALT which implies that ALT is one of the most useful routine test predictors of HCV infection. The finding for ALT coincides with current medical practice [30], as cases with elevated ALT levels are more likely to be infected with HCV. However, ALT elevation is not specific for viral hepatitis [31] and so an increased level of ALT in the blood is used as a guide to request specific second-tier tests that may include HepC immunoassay.

RDW has also been previously identified as having a relationship with HCV in a small-scale Chinese in-patient study [18]; the current analysis supports that finding in a much larger Australian community cohort. Lymphocytes have also been identified as playing a role in HCV infection in a systematic review of epidemiological studies [32].

Age is another strong predictor of HCV status, with increasing age decreasing the chance of being HCV positive. Evidence from Europe [33] and Australia [34] supports this. The increased prevalence in this study of HCV diagnosis in younger patients suggests that clinicians could use this evidence to be proactive in seeking information on the key variables identified in this study in the younger age groups. Although the laboratory supplying the data is housed on a hospital campus, the organisation provides pathology services to the public as well as inpatients, hence the sample contains many community-living persons with a wide range of clinical indications leading to the request for an HCV immunoassay.

The multiple imputation methods employed here assume MAR. However, the results of this study are limited by the likelihood that MNAR is actually the type of missingness observed in pathology data. Policy initiatives such as those recently employed for Vitamin D [35] provide external reasons to reduce the number of pathology tests ordered in a variety of situations. Another likely reason that MNAR would arise in the context of this study is that

clinicians are likely to use their clinical judgement to order some tests and not others. This clinical judgment is now also being enhanced by data mining [36]. Non-clinical characteristics can also drive choice of tests requested [37]. Practice characteristics and other clinical notes are not captured in the available data.

A second and related limitation of this multiple imputation analysis is that were MAR to be assumed, the parameter estimates from the imputed data are very sensitive to the model employed for the probability of response [13]. The probability of response can be modelled by using linear regression models for continuous variables, logistic regression models for binary variables, and multinomial logistic regression models for categorical variables with more than two classes.

A final limitation of this analysis concerns the rarity of the outcome: only 3% of subjects tested positive for HCV. Whilst the rarity of the outcome supports our view that inclusion bias is likely to be small, it does raise issues around the stability of classical estimation algorithms. Logistic regression for rare events may require the use of penalised likelihood instead of maximum likelihood [38].

5 Conclusion

Faced with missing data in a routine pathology laboratory database, logistic regression to develop a prediction model for HCV positive assay results was successfully undertaken following the use of multiple imputation. The coefficient estimates of the regression models built on various imputed datasets were pooled to obtain overall estimates. These logistic regression models, on average, explained about 70% of the variation in the imputed dataset. The regression models built on imputed datasets have similar AUROCs compared to the regression model generated using the complete dataset (73% on complete dataset compared to an average of 71% to 72% on multiply imputed datasets). However, the regression models on the imputed datasets identified more predictors as significant in predicting HCV infection. This means that MI leads to richer prediction models and not using MI may lead to missing predictors that have a useful relationship with the outcome of interest. This study suggests that MI in the diagnostic process of HepC infection using lab-

oratory data has identified combinations of routinely measured biomarkers that are integral to the prediction of HCV infection.

Acknowledgements: The authors wish to thank Dr Gus Koerbin and staff at ACT Pathology for their support of this project. They also thank Mr Sheikh Faisal who undertook the initial analysis of this data as part of his Master's degree research.

Financial support: This research did not receive any specific grant from funding agencies in the public, commercail, or not-for-profit sectors.

6 References

1. Richardson AM, Signor BM, Lidbury BA, Badrick T. Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. Clin Biochem 2016; 49: 1213-20. http://dx.doi.org/10.1016/j.clinbiochem.2016.07.013.

2. Richardson AM, Lidbury BA. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. BMC Bioinformatics 2013; 14: 206-13. doi: 10.1186/1471-2105-14-206.

3. Richardson AM, Lidbury BA. Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines. BMC Med Inform Decis Mak 2017; 17: 121. DOI 10.1186/s12911-017-0522-5.

4. Rubright JD, Nandakumar R, Glutting JJ. A simulation study of missing data with multiple missing X's. Practical Assessment, Research and Evaluation 2014; 19: 1-8.

5. Stanaway JD Flaxman AD, Naghvi M, Fitzmaurice C, Vos T, et al. The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. Lancet 2016; 388: 1081-88. https://doi.org/10.1016/S0140-6736(16)30579-7

6. Polaris Observatory HCV Collaborators. Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study. Lancet Gastroenterol Hepatol 2017; 2: 161 – 176. doi: 10.1016/S2468-1253(16)30181-9

7. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology 2012; 142: 1264-73.doi: 10.1053/j.gastro.2011.12.061.

8. Razali K et al. Modelling the hepatitis C virus epidemic in Australia. Drug Alcohol Depend 2007; 91: 228-35. DOI:10.1016/j.drugalcdep.2007.05.026.

9. Vietri J, Prajapati G, El Khoury AC. The burden of hepatitis C in Europe from the patients' perspective: a survey in 5 countries. BMC Gastroenterol 2013; 13: 16. doi: 10.1186/1471-230X-13-16.

10. WHO. Guidelines for the care and treatment of persons diagnosed with chronic Hepatitis C virus infection. Geneva: WHO; 2018.

11. European Association for the Study of the Liver. EASL clinical practice guidelines: management of hepatitis C virus infection. J Hepatol 2014; 60: 392 – 420. doi: 10.1016/j.jhep.2013.11.003

12. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am Stat 2001; 55: 244-54. https://doi.org/10.1198/000313001317098266

13. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New York: Wiley; 2002.

14. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B Stat Methodol 1977; 39: 1-38. http://dx.doi.org/10.2307/2984875

15. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

16. Janssen KJM, Vergouwe Y, Donders RT, Harrell FE, Chen Q, Grobbee DE, Moon KGM. Dealing with missing predictor values when applying clinical prediction models. Clinical Chemistry 2009: 55; 994 – 1001. DOI: 10/1373/clinchem.2008.115345.

17. Waljee AK, Mukherjee A, Singal AG, et al Comparison of imputation methods for missing laboratory data in medicine BMJ Open 2013;3:e002847. doi: 10.1136/bmjopen-2013-002847

18. Shang G, Richardson AM, Gahan ME, Easteal S, Lidbury BA. Predicting the presence of Hepatitis B virus surface antigen in Chinese patients by pathology data mining. J Med Virol 2013; 85: 1334–39. DOI 10.1002/jmv.23609.

19. Van Buuren S. Flexible imputation of missing data, 2nd edition. Boca Raton FL: CRC Press; 2012.

20. Yang, X. , Belin, T. R. and Boscardin, W. J. (2005), Imputation and Variable Selection in Linear Regression Models with Missing Covariates. Biometrics, 61: 498-506. doi:10.1111/j.1541-0420.2005.00317.x

21. Su YS, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. J Stat Softw 2011; 45: 1-31.doi:10.18637/jss.v045.i02

22. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.

23. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011; 45: 1-67. http://dx.doi.org/10.18637/jss.v045.i03

24. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005; 21(20): 7881.DOI: https://doi.org/10.1093/bioinformatics/bti623

25. Gross J, Ligges U. nortest: Tests for Normality. R package version 1: 0-4; 2015.

26. Revelle W. psych: Procedures for Personality and Psychological Research. Iliinois: Northwestern University; 2018.

27. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.

28. Von Hippel PT. Regression with missing Ys: an improved strategy for analysing multiple imputed data. Sociological Methodology 2007; 37: 83 - 117. doi:https://doi.org/10.1111/j.1467-9531.2007.00180.x

29. Lidbury BA. Predicting liver disease post hepatitis virus infection: In silico pathology and pattern recognition. EBioMedicine 2018,35: 10 - 11. doi: 10.1016/j.ebiom.2018.08.032

30. Holmes J, Thompson A, Bell S. Hepatitis C an update. Aust Fam Physician 2013; 42: 462-66.

31. Kwo PY, Cohen SM, Lim JK. ACG clinical guideline: evaluation of abnormal liver chemistries. Am J Gastroenterol 2017; 112: 18-35. DOI: 10.1038/ajg.2016.51.

32. He Q, He Q, Qin X, Lei S, Li T, Xie L, Deng Y, He Y, Chen Y, Wei Z. The relationship between inflammatory marker levels and hepatitis C virus severity. Gastroenterol Res Pract 2016; 2978479. http://dx.doi.org/10.1155/2016/2978479.

33. Cacoub P, Comarmond C, Domont F, Savey L, Desbois AC, Saadoun D. Extrahepatic manifestations of chronic hepatitis C virus infection. Ther Adv Infect Dis 2016; 3: 3 -14. https://doi.org/10.1002/1529-0131(199910)42:10%3C2204::AID-ANR24%3E3.0.CO;2-D.

34. Faustini A, Colais P, Fabrizi E, Bargagli AM, Davoli M, et al. Hepatic and extra-hepatic sequelae, and prevalence of viral hepatitis C infection estimated from routine data in at-risk groups. BMC Infect. Dis 2010; 10: 97. https://doi.org/10.1186/1471-2334-10-97.

35. Boyages SC. Vitamin D testing: new targeted guidelines stem the overtesting tide. Med J Aust 2016; 204: 18. DOI: 10.5694/mja15.00497

36. Lidbury BA, Richardson AM, Badrick T. Assessment of machine-learning techniques on large pathology data sets to address assay redundancy in routine liver function test profiles. Diagnosis 2015; 2: 41 – 51. DOI: 10.1515/dx-2014-0063.

37. Smellie WSA, Galloway MJ, Gedling P. Is clinical practice variability the major reason for differences in pathology requesting patterns in general practice? J Clin Pathol 2002; 55: 12 – 314. https://doi.org/10.1136/jcp.55.4.312

38. King G, Zeng L. Logistic regression in rare events data. Political Analysis 2001; 9: 137-163. https://doi.org/10.1093/oxfordjournals.pan.a004868.

Chapter 8

Conclusion

Multilevel modelling is a widely used statistical method for analysing clustered data. Such datasets allow for every individual to have its own profile, with measurements captured at each level in the dataset. The natural structure of such models makes the consequences of discarding observations with missing data more severe. The importance and need to account for missing values in such data structures increases as the levels in the framework increase. Handling missing data at Level 2 has received less attention in the literature, with even fewer studies focusing on the impact of missing values occurring in more than two levels. Recent literature has seen several calls to extend the current methodology of MI to handle data structures with more than two levels (van Buuren, 2018; Lüdtke et al., 2017; Gibson & Olejnik, 2003).

The purpose of this thesis was to extend and examine the performance of two missing data approaches for multilevel models; multiple imputation using chained equations (MICE) and Joint Modelling (JoMo), against Complete Case Analysis (CC). This study developed the two extensions, MICE (ext.) and JoMo (ext.) as described in sections 4.4 and 4.5 respectively, and used a simulation to compare the multiple imputation extensions with Complete Case Analysis under circumstances known to impact both missing data methods and hierarchical modelling. This thesis simulated multilevel data, and imposed an array of conditions such as mechanism of missingness, proportion of missing data, and location of missingness; on covariates measured at each level in the data. The datasets were then submitted to all three MI methods for comparison in their ability to retrieve parameter estimates close to their true values and obtain unbiased estimates.

My original contribution to knowledge is enabling R software to extend the current practices of imputation, MICE (ext.) and JoMo (ext.) (see sections 4.4.1 and 4.5.1) to handle missing values in covariates measured up to the third level. Details of how this was done has been provided in **Chapter 4**, and the corresponding codes have been provided in the Appendix. In this journey of extending the current MI practices, a few gaps in the literature surrounding multilevel multiple imputation were revealed. Researchers in this field have restricted their study to imposing missing values in either level 1 variables and their cluster means (Grund, Lüdtke, & Robitzsch, 2018; Resche-Rigon & White, 2018), or at level 1 and level 2 individually (Medhanie, 2013; van Buuren, 2018). This thesis is among the first few to consider imputation of partially observed covariates across multiple levels in the dataset. Imputation of missing data observed in a single variable or multiple variables at level 1, is straightforward. However, things get more complicated when missing values occur in multiple variables measured at multiple levels. Since such situations frequently occur in research, we need sophisticated methods and researchers trained in applying these methods to impute them properly. This thesis provides extensions to two popular imputation methods, namely MICE and JoMo, that facilitates this, with a tutorial style chapter (see Chapter 4) that demonstrates to users how this can be done in practice. In addition, the chapter also provides the procedures on how to extend MICE (see section 4.4.1) and JoMo (see section 4.5.1). The final code used for simulation is provided in the Appendix.

In this thesis we demonstrated the impact of varying the number of clusters at each level in the dataset on the parameter estimates obtained from implementing MICE(ext.) and JoMo(ext.), and an adjustment for bias in the confidence intervals to improve coverage probabilities. This thesis is among the first to apply the the adjustments for confidence intervals as proposed by Menéndez et al. (2014) in practice especially within the MI framework.

There is little to no literature studying the impact of varying cluster sizes on MI, especially in the context of three level datasets. This thesis concluded that the exten-

sions of MICE for three level datasets were robust across varying number of clusters at each level, with low levels of ARB and MSE.

Chapter 6 also provides a solution to the dilemma of the low coverage of confidence intervals obtained. Finally, in **Chapter 7**, we demonstrated a real world application of this method using the National Family Health Survey data to show impact of multilevel multiple imputation in the context of maternal and child anaemia in Northern states of India.

The conclusion is divided into four sections. The first section discusses the results obtained through this research and elaborates on some explanations for some unexpected results observed in the study. The second section provides recommendations to medical and public health researchers on how to analyse cross sectional data with multiple levels; this is distinct from longitudinal data with repeated measures; and with missing values. The third section states the limitations of the study. Finally, the fourth section provides recommendations for future research in the area of missing data in multilevel data structures.

8.1 The Results from this thesis

In this thesis, we provided extensions within two popular methods, namely MICE and JoMo, to impute for multilevel missing data in the third level, and compared their performance. It was positive to note the extensions performed well, with low values of relative bias and mean squared errors. This has been discussed in detail in section 5.4. In this section, we shine light on key results that were obtained in the course of this study.

This thesis successfully attempted to fill in some of the gaps that were highlighted in Figure 3.4.2. This translated into the following action items.

- Extending the current theory and practical implementation of MI to handle missing values in the third level.
- Testing the performance of MI for missingness occurring together across mul-

tiple levels in the dataset.

- Testing the performance of extensions of MI methods to handle derived variables in the context of multilevel models.

In **Chapter 4**, we demonstrated how we applied and measured the performance of MI separately for every level 2 unit as proposed by Gelman and Hill (2006), and extended the method to level 3 units using MICE and JoMo in R. We introduced varying proportions of missing values in covariates captured at all three levels in the dataset. In addition, derived variables such as cluster means and deviations were included in both the analysis and imputation models. Adding these "aggregates" or "contextual variables" allowed us to estimate both within and between group effects. This was demonstrated by including the deviation of variables from their cluster means and the cluster means themselves in the imputation and analysis model. This allowed us to estimate how far the variables were from their cluster means and also allowed us to trace any effect of imputations on this relation. Performing the imputations separately at each level, allowed us to preserve the multilevel structure. We were able to define the nesting for every variable in the model, and the clustered hierarchical nature of the higher level variables that were to be imputed. This feature allows users to extend the imputation strategy for handling missing values in more that three levels. The results obtained from the simulation study are elaborated on in **Chapter 5**. It is important to note that the number of individuals/household, households/PSU and PSUs are chosen to maintain the observed structure. True household sizes have been substituted for a fixed value of a typical household size observed in the NFHS dataset. Given that cluster-level means are being used, then the variance of this variable depends on the size of cluster even if the distribution is identical at level 1, so any method using the cluster-level means would be affected with unbalanced cluster sizes.

In **Chapter 5**, we present the results of our simulation study after extending both MICE and JoMo to impute for missing values in the third level. We observed low values of bias and MSE, and high probability coverage of CIs for both MICE(ext.)

and JoMo (ext.) when MCAR was observed at level 1 alone. Coverage Probabilities deteriorated when missing values were introduced in higher levels. This was especially observed in derived variables, and alternate methods of imputation for derived variables were explored. A more extreme scenario was explored, where we increased the proportion of missingess at level 1 from 20% to 50%. The MI extensions performed considerably better than Complete Case Analysis, with low bias and MSE values, and greater probability coverage of CIs for the parameter estimates.

In **Chapter 6**, we addressed the robustness of the MI methods in handling varying number of clusters at each level. In the previous chapters, we had kept the number of clusters across all three levels consistent in all our analysis, and varied the proportion and mechanisms of missingness observed. In Chapter 6, we tested how the results change as the number of units at each level in the dataset vary. To achieve this, we studied various combinations of the initial set-up i.e. four units at level 1, nested within 20 units at level 2 that are further nested within 100 units at level 3, or $4 \times 20 \times 100$. In addition, we also addressed the non-coverage that was observed in the extensions of MICE and JoMo (see Tables 5.4.4 and 5.4.3), and demonstrated higher coverage of confidence intervals post adjustment for bias (see Figures 6.4.4 and 6.4.5).

Chapter 7 provided an illustration of each component in the thesis i.e. multiple imputation, multilevel modelling and the intersection that it at the core of the thesis, multilevel multiple imputation. This chapter had three case studies each addressing one of the components. This chapter was important as it allowed the reader to understand the practical implementation of the methods discussed beyond the simulation set up and the robustness of MI methods to noisy datasets. The first case study was an illustration of imputation of partially observed covariates in a three level dataset. The study employed a sequential process to handle missing data by dividing the datasets based on whether maternal and child haemoglobin levels were observed or not. The study followed this by classifying both mothers and child as anaemic and non-anaemic after imputing all missing values in the dataset. The study provided

great insight into the effect of individual, maternal and household risk factors on prevalence of anaemia among children under 5 years in India - a key finding being the positive impact of antenatal visits on the child's Hb levels, especially among non-anaemic women. The second case study emphasised the impact of incorporating the structure of the dataset into analysis and the implications of ignoring the sampling design in analysis. This was demonstrated using the third National Family Health Survey to identify a risk profile for HIV in the Indian population. Finally, the final case study showed the illustration of the novel approach of using influx and outflux statistics to improve prediction of Hepatitis C using pathology variables in the presence of missing data. This paper is among the first to illustrate the use of these influx and outflux statistics (as proposed by van Buuren(2018)) in practice, especially in the context of imputation of partially observed pathology data.

8.1.1 Reflections on the Results obtained

Previous research suggested that MI would come out as a stand-out winner (Little & Rubin, 2014; Lee & Carlin, 2010; Liu & De, 2015) as opposed to Complete Case Analysis, though it continues to remain prevalent in literature (Karahalios et al., 2012). While the estimates obtained through MICE and JoMo showed low MSE and ARB values, Complete Case Analysis performed similarly to the MI methods. The only exception was in estimating the intercept, which produced very large MSE values when Complete Case Analysis was employed (see Table 5.4.4). These results were particularly surprising as one would expect a worse performance by Complete Case Analysis as the missing values were being introduced in higher level units. This suggested that the gains in obtaining good estimates by performing MI were marginal when compared to Complete Case Analysis. Surprisingly, the standard errors of the estimates (with the exception of the intercept) are only slightly smaller for the three level extensions of MICE and JoMo, compared to Complete Case Analysis.

The method proposed in this thesis is conceptually similar to the method proposed by Bartlett, Seaman, White, Carpenter, and Initiative* (2015) in that in both the meth-

ods, compatibility between the imputation and substantive models is avoided by specifying a joint model for outcome and covariates based on which the conditional distribution of outcome given covariates matches the substantive model and then using the imputation model implied by this joint model. As with the SMC-FCS, the extensions proposed in this thesis also recommend performing MI on the covariates followed by imputation of the derived variables.

The extensions of MICE and JoMo proposed in this research allow researchers to specify the structure and nesting of variables up to three levels in the dataset. This allows users to specify an imputation model for each variable at each level separately, and preserve the multilevel structure of the dataset. Using this technique we can define the hierarchical clustering present for every variable in the dataset. In reflection, this may create a limitation for users who want to impute for datasets that have multiple membership or cross classification in them, as it would be difficult to define the on hierarchical structure for variables in such a dataset. This is because the degree to which each lower level unit belongs to a higher level unit varies across the dataset, posing new challenges to researchers.

We also added aggregates or contextual variables that are referred to as derived variables in this thesis, to estimate between-group effects (cluster means) and within-group effects (deviations from cluster means). As indicated by Enders (2008), including auxiliary variables in the imputation and analysis model proved beneficial. An anomaly observed in the coverage that was discussed in detail in the thesis (see section 5.6.3), was the 0% probability coverage of confidence intervals for the derived variables, compared to the almost 100% probability of coverage of other variables in the model. We first hypothesized that this might be due to a shift in the distribution of the estimates obtained due to the passive imputation approach that was taken for imputation of derived variables. Referring back to 0% coverage of confidence intervals that was discussed again in Chapter 6, we see an improvement in the coverage probabilities after the interval estimates were adjusted for bias.

The credible interval plots (see Figure 5.4.4) increased our confidence that the intervals produced by both MICE (ext.) and JoMo (ext.) captured the true values of the parameters (except in Scenario 3 which is an extreme scenario). In section 5.6.3, we compared alternate approaches like the "Just Another Variable" approach to handle missing values in the derived variables in the model.

There remains an ongoing discussion among applied researchers on handling missing values for derived variables and also on imputing categorical covariates derived from an underlying continuous variable. MI methods for categorical data continues to remain a challenging task, as estimation issues increase in the the imputation model, as the number of parameters increases with increased number of categories and variables. Ideally, researchers would prefer imputation models that preserve the relationship between variables while requiring a moderate number of parameters to estimate (Audigier, Husson, & Josse, 2017).

While our study established passive imputation as the superior method for handling derived variables in comparison to the JAV, passive imputation continues to pose a quandary to applied researchers. Researchers remain uncertain about imputing derived variables (eg: BMI) as is, or imputing the underlying variables (height and weight) and then recalculating the derived variable (eg: BMI = weight/height^2). In his study, von Hippel (2009) proposed the method of **Passive Imputation** and proved its superiority to other methods in handling non-linear transformations such as squares and interactions in datasets, however its performance in handling derived variables in the multilevel context was unexplored. Our comparison with alternate methods, such as the JAV approach to handle derived variables (see section 5.6.3) confirms Passive Imputation as the better method to handle derived variables such as cluster means and deviations in the context of multilevel data.

8.1.2 Reflections from the Case Studies

The three case studies in Chapter 7 provide novel insights into the application of the methods explained in the simulation studies in Chapters 5 and 6. All the case studies not only provide assurance of applicability of the statistical methods discussed in the leading chapters, but also provides novel insights into the biological and public health aspects studied in each paper. In this section, we review and reflect on the biological or public health conclusions of each case study.

Case Study 1: Application of three level multiple imputation in national surveys.

Anaemia is a condition characterised by a reduction in the number of red blood cells produced by the body and has been known to impair health and well-being in adults and children worldwide.Morbidity, mortality and economic ramification associated with anaemia are numerous and have a lifelong impact. In children, anaemia may contribute to poor growth, stunting, developmental cognitive delay and increased susceptibility to infections (Swaminathan et al., 2019). Markets are failing to address the issues around malnutrition especially among families that cannot afford to buy adequate food or healthcare (Shekar, Heaver, Lee, et al., 2006). The causes of anaemia are multifaceted and is influenced by a combination of biological, social and environmental factors. Improving nutrition by providing reliable estimates of these indicators is important to monitor national progress towards reducing health inequalities.

In this case study, we performed segmented stepwise MI, where the imputed dataset at each step contributed the prior information to build an imputation model for each successive analysis. Each iteration becomes more comprehensive compared to Complete Case Analysis as we include more observations with partially observed covariates measured at each level, thus reflecting the complex health scenario in India. The results of the imputation and consequent analysis showed a high accuracy (98%) in classification of both mothers and children as anaemic or non-anaemic after imputation of missing haemoglobin levels. Imputing higher level variables also allowed us

to see the impact of biological, maternal and environmental variables on the child's haemoglobin levels. A key finding in the study was the positive impact of increased antenatal visits on the child's haemoglobin levels. This was specifically observed in groups where women were observed as non-anaemic. Despite this observed positive impact, we noted that under 5% of women in northern states of India made more than 10 antenatal visits. Family structure, and mother's education were identified as critical factors influencing the anaemia classification of the child. Contradictory to existing literature, we observed that households with higher wealth index were predicted to have children with lower haemoglobin levels compared to middle class households. While an anomaly, this result further cemented our understanding of anaemia being a multifaceted issue that cannot be boxed as affecting only poorer households of north India.

Case Study 2: Impact of accounting for data structure in national surveys.

The objective of this case study was to highlight the importance of incorporating the structure of the data into analysis and the implications of not doing so. The public health issue highlighted in this study was the prevalence of HIV in India. Univariate models were used to identify the variables to retain in the final model. The unadjusted odds ratios (ORs) were calculated. Finally, we used the intra cluster correlation coefficient (ICC, denoted by ρ) to assess the relatedness of each pair of variables within the clusters at both Primary Sampling Unit (PSU) and household level. A logistic regression model for the survey data was then fit using the same variables as were included in the final model, without accounting for the multiple levels in the dataset.

Although several studies have addressed the risk factors associated with HIV in the Indian population, the importance of accounting for sampling strategy in the analysis has not been completely recognized. We observed a significant variation in the risk estimates obtained when the multilevel structure of the data was included in the analysis. No knowledge of contraceptive methods, age at first marriage, wealth

index and non-coverage of PSUs by anganwadis were significant risk factors for HIV when the multilevel model was used. Blood transfusion is proved a risk factor for transmission of HIV; whereas the single level analysis reported ever having had a blood transfusion as a protective factor against HIV. This was a significant irregularity obtained in our results. Blood transfusions could be a good screening tool, thus reducing the odds of transmission of HIV via blood. Finally, we identified that the low perceived levels of awareness of methods to reduce transmission of HIV served as a major threat to India's initiative to curb the spread of HIV, and demanded serious attention of public health officials.

Case Study 3: Developing a rich imputation model for imputing laboratory data

In this study we provided an analytical investigation of the laboratory diagnosis of Hepatitis C (HCV) infection and focused on how to maximize the predictive value of routine pathology data. This study is among the first to illustrate the practical application of the Influx-Outflux values to help develop an imputation model and identify potentially powerful predictors of missing values. Complete Case Analysis and Multiple Imputation using Chained Equations were used to accommodate missing values in the dataset. Logistic regression model and stepwise selection method were used for analysing the imputed datasets.

The study identified over 100 patterns of missingness in the predictor variables across 14,320 community-patients aged 15 - 100 years accessed via ACT Pathology, The Canberra Hospital, Australia. Biomarkers with around 15% missingness included sodium, potassium, creatinine and urea which are routinely requested for most laboratory investigations and are associated with the Urea-Electrolytes-Creatinine battery of tests. Other biomarkers such as ALB, ALT, ALP, GGT and Total Bilirubin are primary markers sought when HCV is suspected. The study results showed that increased levels of ALT, RDW and lymphocytes increased the odds of detecting HCV infection. RDW was also identified as having an association with HCV in the Australian context. The high prevalence of HCV diagnosis among younger age groups

is sufficient evidence for clinicians to be proactive in seeking information on these key biomarkers in the study. The study demonstrated that implementation of MI to handle partially observed covariates can result in richer prediction models, while not using MI may result in exclusion of potentially powerful predictors that are have clinically significant relationship with the outcome of interest.

8.2 Recommendations in Practice for Applied Researchers

8.2.1 Prior to Collecting Data

The thesis highlights the need to minimise missing data in research. As mentioned earlier, prevention is the best strategy to handle missing data. Unfortunately, this is easier said than done. No missing values are extremely unlikely especially in research that is conducted in an uncontrolled setting. Thus the researcher should start thinking about the probability of occurrence of missing data even before data in captured. The thesis has illustrated that as the proportion of missingness increases, the more severe is the consequence. Even after implementing a sophisticated approach of multiple imputation that preserved the structure of the data and accounts for the relationship between variables at each level, we observe that the performance of the MI methods decreases with increased proportions of missingness. Thus, the best way to minimise this effect, is to take necessary steps prior to data collection to reduce this impact.

This may require the researcher to take a step back and identify topics that are sensitive that participants may not be comfortable in providing. It would also help the researcher identify certain groups that may not be comfortable answering these questions. Non-response poses the greatest threat to surveys, and the patterns of missingness observed often suggest that the achieved samples may not be a true representation of the population. If a survey is created in a way that it results in a subset of the population refusing to participate in the study, it could result in non-response bias. Non-response should be accounted for in sample size calculation, and non-response weights can be used to adjust for this in surveys. Inaccessibility of ar-

eas during data collection can also result certain areas being excluded from analysis. This should be noted and reported accordingly.

Another common reason for missing data in surveys is due to a large number of skip questions, as a result of which participants do not answer certain questions in the questionnaire. Skip questions can be classified as MAR. If a person says they don't smoke, then the number of cigarettes smoked per day will be missing, but this is missing dependent on observed data. One of the basic assumptions of MI in general is that the data should be MAR. This continues to be an assumption even when considering extending MI to three level data structures.

The issue is that this assumption is violated quite easily. Even when the MAR assumption is satisfied, careful consideration is needed when developing the imputation model to ensure that the analysis model is congenial with the imputation model (see Section 2.3.8). Assuring confidentiality of the information captured and distributed can also help increase participation in the study, and encourage study participants to be more confident in sharing their information. The imputation methods in this thesis require the assumption of MAR to be satisfied. One of the biggest challenges in the field, is that without making distributional assumptions (themselves untestable) you cannot distinguish between MAR and MNAR. However, if researchers suspect that the data might be missing not at random, it is possible to use the methods mentioned here with caution, and by including additional dummy variables as missingness indicators in the imputation model. However, currently this approach is still highly conceptual in the literature, and further research is necessary to confirm how one might apply this in practice.

8.2.2 Upon Data Collection and in Analysis

It is important that researchers study the data to identify **missing data patterns**. Continuous monitoring of the data can also help flag any frequent occurrences of missingness in the data and help identify queries to raise with the field investigators

collecting the data. It is important that values be coded correctly. For example, values pertaining to *"Don't Know"* or *"Not Applicable"* should not be coded the same as *"Missing"*. In practice, missing values are often coded as 9/99/999/9999 for numeric variables or "." for string variables. It is important to recheck that these values are not lost when the dataset is imported into statistical software.

It is also important to understand the **proportion of missingness** in the dataset, and the variables in which missingness occurs. Researchers can make use of Matrix plots and other visualisation tools mentioned in section 4.1. This would provide an understanding of any obvious patterns of missingness in the datasets, or clusters of variables that are missing together. Researchers can make use of the **influx and outflux statistics** (see Section 2.3.7) to understand the variables that would be useful in constructing an imputation model, and variables that are associated with the variable exhibiting missingness. This can be further visualised using the fluxplot. An illustration of this approach has been provided in Case Study 3 (see section 7.4).

With this information, the researcher can now proceed to develop the imputation model. Using the imputation algorithm specified by van Buuren (2018), for a two level model (see Table 4.1.1) or the algorithm developed in this thesis for a three level model (see Table 4.4.4); the researcher can proceed with imputing any partially observed covariates. The researcher should make sure that the imputation model is congenial with the analysis model, and includes all predictors, transformations or interactions that are specified in the analysis model. If the analysis model includes linear (or non-linear) transformations of a covariate with missing values, our study recommends passive imputation using the imputed value of the variable with missing data. Regardless of the method used to treat missing data, it is important that researchers report the proportion of missingness observed in the study, the reason for the missingness and a detailed description of the method that was employed to handle for missing data. A detailed description of how to implement this approach in software in provided in Chapter 4. The full R code is provided in the Appendix. In addition, an illustration of implementing this algorithm in real world data analysis

has been provided in Case Study 1 in Section 7.2 of the thesis to identify the effect of maternal and household risk factors on anaemia among children.

8.3 Limitations of the Study

This thesis examined the performance of MI under circumstances that do not represent all possible situations when one can employ this technique. The limitations of the thesis are addressed in this section. First, the thesis does not consider longitudinal or non-hierarchical datasets. While these are both important and interesting data structures, this study concentrated on extending the vertical hierarchy to three levels. More information on multiple imputation for longitudinal studies can be found in the research by De Silva, Moreno-Betancur, De Livera, Lee, and Simpson (2017). Second, the simulation assumed fixed ICC values at both level 2 ($ICC_{level2:level3}$ = 0.35) and level 3 (ICC_{level3} = 0.33). These values were chosen as they are in the mid range, and were not the focus of the study. Several studies have already established effect of varying ICC values on multiple imputation, with MICE and JoMo introducing bias in regression coefficients associated with level 2 variables, in conditions with large ICC values, (Grund, Lüdtke, & Robitzsch, 2016; Grund et al., 2018). Grund et al. (2018) stated that single level MI should be avoided unless there is under 5% missingness observed and ICC values are less than 0.10. Third, in order to understand the effect of missing data in an uncluttered three level model, the simulation set-up was restricted such that we imposed missing values in only one variable measured at each level in the dataset. In reality we are surrounded by missing values, and these can potentially occur in every variable captured in the study. Finally, conclusions from the simulation study were restricted to the range of the simulation conditions. The objective of the simulation study was to compare performance of MI across varying proportions, mechanisms and locations of missingness. In doing so we tested only one set of true values of the parameter throughout this thesis. These values were chosen to be far enough away from zero to avoid boundary issues in estimation, as well as to be different from each other for identifiability. We acknowledge that different selections of these values may result in different conclusions.

8.4 Potential Pathways for Future Research

Finally, while this thesis has successfully answered many questions that were posed at the start of this research project, it also provides several avenues to pursue in the future. Implementation of MI, and imputation of variables in survival analysis remains an open question. A more detailed analysis of the impact of varying cluster sizes and number of clusters on MI can provide more insight into the robustness of the extensions developed to the existing methods. As discussed in Section 1.2.1, the results of this thesis cannot be extrapolated to longitudinal datasets and non-hierarchical data structures. It would be interesting to understand the performance of MI in studies with multiple memberships and cross classification within levels. The implementation of passive imputation in handling between level interactions would be pertinent in such scenarios. This thesis is among the first few to illustrate both practical application of the influx-outflux approach, and also interval adjustment to improve nominal coverage probabilities. A more detailed analysis of these methods could further cement their application in practice when constructing rich imputation models.

References

Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57*(3), 273–291. (cited on page 59)

Allison, P. (2001). *Missing data. quantitative applications in the social sciences.* Sage, Thousand Oaks, CA. (cited on page 1)

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal*, *53*(1), 57–74. (cited on pages 22, 57, and 58)

Audigier, V., Husson, F., & Josse, J. (2017). Mimca: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and computing*, *27*(2), 501–518. (cited on page 206)

Audigier, V., & Resche-Rigon, M. (2017). micemd: Multiple imputation by chained equations with multilevel data. *R package version*, *1*(0). (cited on pages 54 and 71)

Audigier, V., White, I. R., Jolani, S., Debray, T. P., Quartagno, M., Carpenter, J., ... others (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, *33*(2), 160–183. (cited on page 5)

Austin, P. C., & Leckie, G. (2018). The effect of number of clusters and cluster size on statistical power and type i error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation*, *88*(16), 3151–3163. (cited on page 120)

Bartlett, J. W., Harel, O., & Carpenter, J. R. (2015). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American journal of epidemiology*, *182*(8), 730–736. (cited on pages 16 and 25)

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., & Initiative*, A. D. N. (2015). Multiple imputation of covariates by fully conditional specification:

Accommodating the substantive model. *Statistical methods in medical research*, *24*(4), 462–487. (cited on page 204)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. (cited on page 94)

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651–675. (cited on page 34)

Bondarenko, I., & Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in medicine*, *35*(17), 3007–3020. (cited on pages 39 and 59)

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (mmmc) models. *Statistical Modelling*, *1*(2), 103–124. (cited on page 58)

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons. (cited on pages 5 and 88)

Chandola, T., Clarke, P., Wiggins, R. D., & Bartley, M. (2005). Who you live with and where you live: setting the context for health using multiple membership multilevel models. *Journal of Epidemiology & Community Health*, *59*(2), 170–175. (cited on page 61)

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. F. (2010). The choice between fixed and random effects models: some considerations for educational research. (cited on page 58)

Cro, S., Morris, T. P., Kenward, M. G., & Carpenter, J. R. (2016). Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. *The Stata Journal*, *16*(2), 443–463. (cited on page 14)

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., & Simpson, J. A. (2017). A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC medical research methodology*, *17*(1), 114. (cited on page 213)

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., & Simpson, J. A.

(2019). Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study. *BMC medical research methodology*, *19*(1), 1–14. (cited on page 14)

Dettori, J. R. (2011). Loss to follow-up. *Evidence-based spine-care journal*, *2*(01), 7–10. (cited on page 14)

Diya, L. (2011). *Multilevel models in health care research: application to nurse staffing research* (Unpublished doctoral dissertation). Doctoral thesis in Bio-medical Science, Leuven. (cited on page 44)

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. (cited on page 64)

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 434–448. (cited on pages 64, 72, and 205)

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological methods*, *19*(1), 39. (cited on pages 49 and 50)

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological methods*, *23*(2), 298. (cited on pages 6, 56, and 57)

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222. (cited on pages 54, 56, 61, 62, and 83)

Fielding, S., Fayers, P. M., & Ramsay, C. R. (2009). Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes*, *7*(1), 57. (cited on page 17)

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press. (cited on pages 5, 20, 21, 44, 45, 47, 54, 76, 77, and 202)

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, *63*(2),

204–238. (cited on page 199)

Gill, J., & Womack, A. J. (2013). The multilevel model framework. *The SAGE handbook of multilevel modeling. London: SAGE Publications Ltd*. (cited on pages 47 and 48)

Goldstein, H., Pan, H., & Bynner, J. (2004). A flexible procedure for analyzing longitudinal event histories using a multilevel model. *Understanding Statistics*, *3*(2), 85–99. (cited on page 5)

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 505–513. (cited on page 61)

Greenland, S. (2000). Principles of multilevel modelling. *International journal of epidemiology*, *29*(1), 158–167. (cited on page 7)

Groves, R. M. (2004). *Survey errors and survey costs* (Vol. 536). John Wiley & Sons. (cited on page 12)

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the r package pan. *SAGE Open*, *6*(4), 2158244016668220. (cited on page 213)

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149. (cited on pages 200 and 213)

Huque, M. H., Moreno-Betancur, M., Quartagno, M., Simpson, J. A., Carlin, J. B., & Lee, K. J. (2019). Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biometrical Journal*. (cited on pages 54 and 56)

Kalaycioglu, O., Copas, A., King, M., & Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(3), 683–706. (cited on page 57)

Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., & Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC medical research methodology*, *12*(1), 96. (cited on pages 2, 3, 11, and 204)

Keogh, R. H., Seaman, S. R., Bartlett, J. W., & Wood, A. M. (2018). Multiple imputation of missing data in nested case-control and case-cohort studies. *Biometrics*, *74*(4), 1438–1449. (cited on page 27)

Kowarik, A., & Templ, M. (2016). Imputation with the r package vim. *Journal of Statistical Software*, *74*(7), 1–16. (cited on page 70)

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage. (cited on pages 45, 76, and 77)

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate behavioral research*, *30*(1), 1–21. (cited on page 49)

Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: comparing joint multivariate normal and conditional approaches. *Political Analysis*, *22*(4). (cited on page 57)

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, *171*(5), 624–632. (cited on pages 29 and 204)

Lee, K. J., & Carlin, J. B. (2017). Multiple imputation in the presence of non-normal data. *Statistics in medicine*, *36*(4), 606–617. (cited on pages 79 and 93)

Lee, K. J., Roberts, G., Doyle, L. W., Anderson, P. J., & Carlin, J. B. (2016). Multiple imputation for missing data in a longitudinal cohort study: a tutorial based on a detailed case study involving imputation of missing outcome data. *International Journal of Social Research Methodology*, *19*(5), 575–591. (cited on page 14)

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons. (cited on pages 23, 57, and 204)

Liu, Y., & De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International journal of statistics in medical research*, *4*(3), 287. (cited on page 204)

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological methods*, *22*(1), 141. (cited on pages 54, 56, 63, 64, 88, and 199)

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data:*

A gentle introduction. Guilford Press. (cited on page 18)

Medhanie, A. G. (2013). The robustness of multilevel multiple imputation for handling missing data in hierarchical linear models. (cited on page 200)

Menéndez, P., Fan, Y., Garthwaite, P. H., & Sisson, S. A. (2014). Simultaneous adjustment of bias and coverage probabilities for confidence intervals. *Computational statistics & data analysis*, *70*, 35–44. (cited on pages 127, 128, and 200)

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538–558. (cited on pages 26 and 38)

Menon, N., Bhaskarapillai, B., & Richardson, A. (2019). Effect of modeling a multilevel structure on the indian population to identify the factors influencing hiv infection. *Biostatistics & Epidemiology*, *3*(1), 126–139. (cited on pages 47, 51, 52, 73, and 74)

Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, *42*(4), 432–466. (cited on page 72)

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102. (cited on pages 95, 102, and 127)

Moschovis, P. P., Wiens, M. O., Arlington, L., Antsygina, O., Hayden, D., Dzik, W., ... Hibberd, P. L. (2018). Individual, maternal and household risk factors for anaemia among young children in sub-saharan africa: a cross-sectional study. *BMJ open*, *8*(5), e019654. (cited on page 52)

Murray, J. S., et al. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, *33*(2), 142–159. (cited on page 35)

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017a). Model checking in multiple imputation: an overview and case study. *Emerging themes in epidemiology*, *14*(1), 8. (cited on pages 19 and 59)

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017b). Model checking in multiple imputation: an overview and case study. *Emerging themes in epidemiology*, *14*(1), 8. (cited on page 59)

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219–242. (cited on pages 16, 27, 29, 56, and 82)

van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC. (cited on pages 2, 16, 18, 21, 24, 25, 33, 37, 59, 62, 63, 66, 68, 69, 70, 72, 82, 84, 87, 114, 115, 118, 120, 172, 199, 200, 204, and 212)

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). Package 'mice'. *Computer software*. (cited on pages 5, 32, 33, 70, and 82)

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological methodology*, *39*(1), 265–291. (cited on pages 33, 40, 41, 60, 114, 115, 118, and 206)

Ono, M., & Miller, H. P. (1969). *Income nonresponses in the current population survey*. US Bureau of the Census. (cited on page 20)

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56. (cited on page 71)

Quartagno, M., & Carpenter, J. (2016). jomo: Multilevel joint modelling multiple imputation.
(cited on pages 5, 6, 31, 71, 88, and 120)

Quartagno, M., & Carpenter, J. (2019b). Package 'jomo'.
(cited on pages 31 and 56)

Quartagno, M., & Carpenter, J. R. (2019a). Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal*, *61*(4), 1003–1019. (cited on page 31)

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from `https://www.R-project.org/` (cited on pages 67 and 68)

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models*. Sage Publications, Thousand Oaks, CA. (cited on pages 46, 49, and 52)

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodol-*

ogy, *32*(2), 143. (cited on pages 36, 39, 40, and 54)

Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical methods in medical research*, *27*(6), 1634–1649. (cited on page 200)

Robitzsch, A., Grund, S., Henke, T., & Robitzsch, M. A. (2019). Package 'miceadds'. (cited on pages 71 and 120)

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. (cited on pages 14, 17, and 94)

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons. (cited on pages 2, 3, 13, 15, 16, 22, 23, 34, and 57)

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, *91*(434), 473–489. (cited on pages 36 and 37)

Rubright, J. D., Nandakumar, R., & Gluttin, J. J. (2014). A simulation study of missing data with multiple missing x's. *Practical Assessment, Research, and Evaluation*, *19*(1), 10. (cited on page 61)

Salim, A., Mackinnon, A., Christensen, H., & Griffiths, K. (2008). Comparison of data analysis strategies for intent-to-treat analysis in pre-test–post-test designs with substantial dropout rates. *Psychiatry research*, *160*(3), 335–345. (cited on page 92)

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press. (cited on page 31)

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147. (cited on pages 5 and 64)

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, *11*(2), 437–457. (cited on page 30)

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, *6*(1), 1. (cited on page 57)

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909–2930. (cited on page 92)

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, *12*(1), 46. (cited on pages 41 and 115)

Shekar, M., Heaver, R., Lee, Y.-K., et al. (2006). Repositioning nutrition as central to development: A strategy for large scale action.
(cited on page 207)

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage. (cited on page 63)

Streiner, D. L. (2008). Missing data and the trouble with locf. *Evidence-based mental health*, *11*(1), 3–5. (cited on pages 20 and 21)

Subramanian, S. V., Kim, D. J., & Kawachi, I. (2002). Social trust and self-rated health in us communities: a multilevel analysis. *Journal of Urban Health*, *79*(1), S21–S34. (cited on page 61)

Swaminathan, A., Kim, R., Xu, Y., Blossom, J. C., Venkatraman, R., Kumar, A., ... others (2019). Burden of child malnutrition in india: A view from parliamentary constituencies. *Economic & Political Weekly*, *54*(2). (cited on page 207)

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2011). Vim: visualization and imputation of missing values. *R package version*, *2*(3). (cited on page 70)

Templ, M., & Filzmoser, P. (2008). Visualization of missing values using the r-package vim. *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology*. (cited on page 68)

Twisk, J. W. (2006). *Applied multilevel analysis: a practical guide for medical researchers*. Cambridge university press. (cited on pages 44 and 54)

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399. (cited on page 35)

Woodbury, M. A., & Orchard, T. (1970). *A missing information principle: theory and applications* (Tech. Rep.). Duke University Medical Center Durham United States. (cited on page 18)

Xie, X., & Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase in-

ference perspective: what happens when god's, imupter' and analyst's models are uncongenial? *Statistica Sinica*, 1485–1545. (cited on page 38)

Yang, X., Belin, T. R., & Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, *61*(2), 498–506. (cited on page 57)

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1874), 2389–2403. (cited on pages 5, 32, 54, 56, 57, and 58)

Chapter 8

Appendix A

The code used to analyse the scenarios from Chapter 5 and 6 provided in the Appendix. As mentioned earlier, $Xijk - tilde$ represents $X_{ijk} - \bar{X}_{jk}$, $Xjkbar - tilde$ represents $\bar{X}_{jk} - \bar{X}_k$, and $Zk - tilde$ represents $Z_{jk} - \bar{Z}_k$. The suffix 20 and 50 denote the degree of missingness in the variables X_{ijk} and Z_{jk}.

Simulation Code for MCAR in Level 1 and Level 2 - MICE

```
rm(list=ls(all=TRUE))
pckgs <- c("sn" ,"lme4" , "devtools" ,"miceadds", "extraDistr", "plyr","lmerTest","mice", "
    mitml", "jomo", "truncnorm")

inst = lapply(pckgs, library, character.only = TRUE) # load them
i = 4 # level1 - individuals
j = 20 # level2 - household
k = 100 # level3 - district
## Population intercept
g_000    <- 2
## Population slope
g_100    <- 2.5  # Xijktilde
g_200    <- 2.5 #Xjkbartilde
g_010    <- 2 #Zktilde
g_020    <- 2.5 #zbark
g_001    <- 3 #wk
#### alpha ####
alpha = c(0.025, 0.975)
true_coef <- c(g_000, g_100, g_200, g_010, g_001)
names <- c("X.Intercept.","Xijktilde","Xjkbartilde","Zktilde", "Wk")
nsims = 1000
simlist = replicate(n = nsims,
                    expr = data.frame(level3 = rep(seq(1:k), each = i*j),
```

```
                                level2 = rep(seq_len(k*j), each = i),
                                level1 = rep(seq(1:i), each = j*k),
                                eijk   = rnorm(n = i*j*k, mean = 0, sd = 1),
                                r0jk   = rep(rnorm(n = k*j, mean = 0, sd = 1), each = i),
                                u00k   = rep(rnorm(n = k, mean = 0, sd = 1), each = i*j),
                                X_ijk  = rsn(n = i*j*k, xi= 70, omega = 20, alpha= 10) ,
                                Z      = rtruncnorm(n = i*j*k, a = 59, b = 396, mean =
                                    114.8, sd = 14),
                                Zjk    = rep(rtpois(j*k, lambda = 3, a = 2, b = 25),
                                    each = i),
                                Z1     = rep(rtpois(j*k, lambda = 5.8, a = 1, b = 41),
                                    each = i),
                                Wk     = rep(rtpois(k, lambda = 3, a = 2, b = 10),  each
                                    = i*j)),
                  simplify = FALSE)
simlist <- lapply(simlist, function(x) within(x, {
  X2 <- (0.6)*Z + sqrt(1-0.6)*Z1
  Z2 <- (0.87)*Zjk + sqrt(1- 0.87)*Z1
  Xbar_jk <- rep(aggregate(X_ijk, list(level2), mean)[,2], each = i)
  Xbar_k <- rep(aggregate(X_ijk, list(level3), mean)[,2], each = i*j)
  Zbar_k <- rep(aggregate(Zjk, list(level3), mean)[,2], each = i*j)
  Xijktilde <- X_ijk - Xbar_jk
  Xjkbartilde <- Xbar_jk - Xbar_k
  Zktilde    <- Zjk - Zbar_k
  Y_ijk <- g_000 + g_100*Xijktilde + g_200*Xjkbartilde +  g_010*Zktilde + g_001*Wk + u00k +
      r0jk + eijk

  #### MCAR in L1 ###########
  Ys <- (Y_ijk - mean(Y_ijk))/sd(Y_ijk)
  X2s <- (X2 - mean(X2))/sd(X2)

  # 20 % missing MCAR
  mcar20x  <- 1- rbinom(i*j*k ,1, 0.20)
  Xijk_20 <- ifelse(mcar20x <1 , NA, X_ijk)
  # 50 % missing MCAR
  mcar50x <- 1- rbinom(i*j*k, 1, 0.50)
  Xijk_50 <- ifelse(mcar50x <1 , NA, X_ijk)

  #### MCAR in L2 ###########
  Yjk <- rep(aggregate(Y_ijk, list(level2), mean)[,2], each = i)
  Y_sdjk <- rep(aggregate(Y_ijk, list(level2), sd)[,2], each = i)
  Y_s <- (Yjk - mean(Yjk))/sd(Yjk)
  Z2s <- (Z2 - mean(Z2))/sd(Z2)
  mcar20z <- rep(1- rbinom(j*k, 1, 0.20), each = i)
```

```
  Zjk_20 <- ifelse(mcar20z <1 , NA, Zjk)
  mcar50z <- rep(1- rbinom(j*k, 1, 0.50), each = i)
  Zjk_50 <- ifelse(mcar50z <1 , NA, Zjk)

  ############ calculate differences for complete case analysis #############
  Xbarjk_20  <- rep(aggregate(Xijk_20, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_20   <- rep(aggregate(Xijk_20, list(level3), mean, na.rm = T)[,2], each = i*j)
  Xbarjk_50  <- rep(aggregate(Xijk_50, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_50   <- rep(aggregate(Xijk_50, list(level3), mean, na.rm = T)[,2], each = i*j)
  X2barjk    <- rep(aggregate(X2, list(level2), mean)[,2],each = i)
  X2bark     <- rep(aggregate(X2, list(level3), mean)[,2], each = i*j)
  Zbark_20   <- rep(aggregate(Zjk_20, list(level3), mean, na.rm = T)[,2],each = i*j)
  Zbark_50   <- rep(aggregate(Zjk_50, list(level3), mean, na.rm = T)[,2],each = i*j)
  Z2bark     <- rep(aggregate(Z2, list(level3), mean)[,2], each = i*j)
  Xijktilde20 <- Xijk_20 - Xbarjk_20
  Xjkbartilde20 <- Xijk_20 - Xbark_20
  Zktilde_20    <- Zjk_20  - Zbark_20
  Xijktilde50 <- Xijk_50 - Xbarjk_50
  Xjkbartilde50 <- Xijk_50 - Xbark_50
  Zktilde_50    <- Zjk_50  - Zbark_50}))

data <- data.frame(simlist[[1]])
sapply(data, function(x) sum(is.na((x))/nrow(data)))

micelist <- simlist
micelist <- lapply(micelist, function(x) { x[c("eijk", "r0jk", "u00k", "Xjkbartilde50", "X_ijk"
     , "Xijktilde50", "Xbark_50",  "Xbarjk_50", "Xijk_50", "mar20x", "X2s","Ys","Xjkbartilde",
     "Xijktilde",  "Xbark_20", "Xbarjk_20")] <- NULL; x })

imp_x <- function(mice_sim){

  variable_levels <- character(ncol(mice_sim))
  names(variable_levels) <- colnames(mice_sim)
  variable_levels[ c("Zjk_20", "Zktilde_20", "Xbar_jk", "Xjkbartilde20", "Yjk", "Z2", "X2barjk"
       ) ] <- "level2"
  variable_levels[ c("Wk", "Xbar_k", "Zbar_k", "X2bark", "Z2bark") ] <- "level3"

  predmat <- make.predictorMatrix(mice_sim)
  predmat[,] <- 0
  predmat[, c("Wk", "level3")] <- -2

  predmat["Xijk_20", c("Y_ijk", "X2", "X2barjk", "Zbar_k", "Wk", "Yjk")] <- c(3, 3, 3, rep(1,3)
       )
  predmat["Zjk_20", c("Z2", "Yjk","Xbar_jk", "Xbar_k", "Z2bark", "Wk")]<-  c(3,3,3,1,1,1)
```

```
  method <- make.method(mice_sim)
  method[c("Xijk_20", "Zjk_20")] <-  "ml.lmer"

  cluster <- list()
  cluster[["Xijk_20"]]<- c("level2", "level3")
  cluster[c("Zjk_20")] <- "level3"
  visit <- c( "Y_ijk", "X2", "X2barjk", "Zbar_k", "Wk", "Yjk", "Xijk_20", "Z2", "Yjk","Xbar_jk"
       , "Xbar_k", "Z2bark", "Wk", "Zjk_20")
  imp_x <- mice(mice_sim, method=method,levels_id=cluster, variables_levels=variable_levels,
       visitSequence = visit,  predictorMatrix=predmat,  seed = 1234)
  impList <- miceadds::mids2datlist(imp_x)
  impList <- within(impList, {
    Xbar_jk <- clusterMeans(Xijk_20, level2)
    Xijktilde20 <- (Xijk_20 - Xbar_jk)
    Xbar_k <- clusterMeans(Xijk_20, level3)
    Xjkbartilde20 <- (Xbar_jk - Xbar_k)
    Zbar_k <- clusterMeans(Zjk_20, level3)
    Zktilde_20   <- (Zjk_20 - Zbar_k)})
  imp20_1 <- datlist2mids(impList, progress = F)

  lm_mice <- with(imp20_1, lme4::lmer(Y_ijk ~ Xijktilde20 + Xjkbartilde20 +  Zktilde_20 + Wk +
        (1|level3/level2)))

  est_mice <- t(summary(pool(lm_mice))$estimate)
  se_mice  <- t(summary(pool(lm_mice))[,2])
  sum_mice  <- c(estimate = est_mice, se = se_mice)
  return(sum_mice)}
imp <- lapply(micelist, imp_x)

mice_sc1 <- data.frame(do.call(rbind, imp))

colnames(mice_sc1)[1:5] <- paste0("est_", names)
colnames(mice_sc1)[6:10] <- paste0("se_", names)
mice_bias_sc1 <- (mice_sc1[1:5] - true_coef)/ true_coef
colnames(mice_bias_sc1) <- names
mice_mse_sc1 <- (true_coef - mice_sc1[1:5])^2
colnames(mice_mse_sc1) <- names

mice_lowerint <- mice_sc1[1:5] + qnorm(alpha[1])*mice_sc1[6:10]
colnames(mice_lowerint) <- paste0("lower_", names)
mice_upperint <- mice_sc1[1:5] + qnorm(alpha[2])*mice_sc1[6:10]
colnames(mice_upperint) <- paste0("upper_", names)
mice_MCse <- mice_sc1[6:10]/sqrt(nsims)
```

```
colnames(mice_MCse) <- paste0("MCse_", names)

mice_sc1_comb <- cbind(mice_sc1, mice_bias_sc1, mice_mse_sc1, mice_lowerint, mice_upperint,
    mice_MCse)

write.csv(mice_sc1_comb, file= "sims_mcar_xz_mice_comb.csv", row.names = F)
```

Simulation Code for MCAR in Level 1 and Level 2 - JoMo

```
i = 4 # level1 - individuals
j = 20 # level2 - household
k = 100 # level3 - district
## Population intercept
g_000    <- 2
## Population slope
g_100    <- 2.5  # Xijktilde
g_200    <- 2.5 #Xjkbartilde
g_010    <- 2 #Zktilde
g_020    <- 2.5 #zbark
g_001    <- 3 #wk

#### alpha ####
alpha = c(0.025, 0.975)
true_coef <- c(g_000, g_100, g_200, g_010, g_001)
names <- c("X.Intercept.","Xijktilde","Xjkbartilde","Zktilde", "Wk")
nsims = 1000
simlist = replicate(n = nsims,
                    expr = data.frame(level3 = rep(seq(1:k), each = i*j),
                                      level2 = rep(seq_len(k*j), each = i),
                                      level1 = rep(seq(1:i), each = j*k),
                                      eijk   = rnorm(n = i*j*k, mean = 0, sd = 1),
                                      r0jk   = rep(rnorm(n = k*j, mean = 0, sd = 1), each = i),
                                      u00k   = rep(rnorm(n = k, mean = 0, sd = 1), each = i*j),
                                      X_ijk  = rsn(n = i*j*k, xi= 70, omega = 20, alpha= 10) ,
                                      Z      = rtruncnorm(n = i*j*k, a = 59, b = 396, mean =
                                          114.8, sd = 14),
                                      Zjk    = rep(rtpois(j*k, lambda = 3, a = 2, b = 25),
                                          each = i),
                                      Z1     = rep(rtpois(j*k, lambda = 5.8, a = 1, b = 41),
                                          each = i),
                                      Wk     = rep(rtpois(k, lambda = 3, a = 2, b = 10),  each
                                          = i*j)),
```

```
                    simplify = FALSE)

simlist <- lapply(simlist, function(x) within(x, {
  X2 <- (0.6)*Z + sqrt(1-0.6)*Z1
  Z2 <- (0.87)*Zjk + sqrt(1- 0.87)*Z1
  Xbar_jk <- rep(aggregate(X_ijk, list(level2), mean)[,2], each = i)
  Xbar_k <- rep(aggregate(X_ijk, list(level3), mean)[,2], each = i*j)
  Zbar_k <- rep(aggregate(Zjk, list(level3), mean)[,2], each = i*j)
  Xijktilde <- X_ijk - Xbar_jk
  Xjkbartilde <- Xbar_jk - Xbar_k
  Zktilde    <- Zjk - Zbar_k
  Y_ijk <- g_000 + g_100*Xijktilde + g_200*Xjkbartilde +  g_010*Zktilde + g_001*Wk + u00k +
       r0jk + eijk

  #### MCAR in L1 ###########
  Ys <- (Y_ijk - mean(Y_ijk))/sd(Y_ijk)
  X2s <- (X2 - mean(X2))/sd(X2)
  # 20 % missing MCAR
  mcar20x  <- 1- rbinom(i*j*k ,1, 0.20)
  Xijk_20 <- ifelse(mcar20x <1 , NA, X_ijk)
  # 50 % missing MCAR
  mcar50x <- 1- rbinom(i*j*k, 1, 0.50)
  Xijk_50 <- ifelse(mcar50x <1 , NA, X_ijk)

  #### MCAR in L2 ###########
  Yjk <- rep(aggregate(Y_ijk, list(level2), mean)[,2], each = i)
  Y_sdjk <- rep(aggregate(Y_ijk, list(level2), sd)[,2], each = i)
  Y_s <- (Yjk - mean(Yjk))/sd(Yjk)
  Z2s <- (Z2 - mean(Z2))/sd(Z2)
  mcar20z <- rep(1- rbinom(j*k, 1, 0.20), each = i)
  Zjk_20 <- ifelse(mcar20z <1 , NA, Zjk)
  mcar50z <- rep(1- rbinom(j*k, 1, 0.50), each = i)
  Zjk_50 <- ifelse(mcar50z <1 , NA, Zjk)

  ############ calculate differences for complete case analysis #############
  Xbarjk_20  <- rep(aggregate(Xijk_20, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_20   <- rep(aggregate(Xijk_20, list(level3), mean, na.rm = T)[,2], each = i*j)
  Xbarjk_50  <- rep(aggregate(Xijk_50, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_50   <- rep(aggregate(Xijk_50, list(level3), mean, na.rm = T)[,2], each = i*j)
  X2barjk    <- rep(aggregate(X2, list(level2), mean)[,2],each = i)
  X2bark     <- rep(aggregate(X2, list(level3), mean)[,2], each = i*j)
  Zbark_20   <- rep(aggregate(Zjk_20, list(level3), mean, na.rm = T)[,2],each = i*j)
  Zbark_50   <- rep(aggregate(Zjk_50, list(level3), mean, na.rm = T)[,2],each = i*j)
  Z2bark     <- rep(aggregate(Z2, list(level3), mean)[,2], each = i*j)
```

```
  Xijktilde20 <- Xijk_20 - Xbarjk_20
  Xjkbartilde20 <- Xijk_20 - Xbark_20
  Zktilde_20    <- Zjk_20  - Zbark_20

  Xijktilde50 <- Xijk_50 - Xbarjk_50
  Xjkbartilde50 <- Xijk_50 - Xbark_50
  Zktilde_50    <- Zjk_50  - Zbark_50}))

data <- data.frame(simlist[[1]])
sapply(data, function(x) sum(is.na((x))/nrow(data)))

jomolist2 <- simlist
jomolist2 <- lapply(jomolist2, function(x) { x[c("eijk" , "r0jk","u00k", "X_ijk", "Z" , "Zjk" ,
     "Z1" , "Ys","X2s", "mar20x","mar50x" ,"Xijk_50" ,"Y_sdjk","Y_s" , "Z2s", "mar20z","mar50z"
      , "Zjk_50","Y_sdk","Y_sk","mar20w","mar50w", "Wk_50","Xbarjk_20", "Xbark_20" , "Xbarjk_50
     ", "Xbark_50", "Zbark_20"  ,"Zbark_50"  , "Xijktilde50" , "Xjkbartilde50" , "Zktilde_50")]
      <- NULL; x })

jomo_x2 <- function(jomo_sim){
  Y = jomo_sim[, c("Xijk_20", "Xijktilde20", "Zjk_20", "Wk")]
  clus = jomo_sim[, c("level2")]
  X = jomo_sim[, c("Y_ijk", "X2", "Xjkbartilde", "Zktilde",  "Zbar_k" , "level2", "level3")]
  impj <-jomo1ran(Y=Y, X = X, clus = clus , nburn=20, nbetween=20, nimp=10)
  jomo_mids2 <- miceadds::jomo2mids(impj)
  jomolist2 <- miceadds::mids2datlist(jomo_mids2)

  jomolist2 <- within(jomolist2, {
    Xbar_jk <- clusterMeans(Xijk_20, level2)
    Xijktilde20 <- (Xijk_20 - Xbar_jk)
    Xbar_k <- clusterMeans(Xijk_20, level3)
    Xjkbartilde20 <- (Xbar_jk - Xbar_k)
    Zbar_k <- clusterMeans(Zjk_20, level3)
    Zktilde_20 <- (Zjk_20 - Zbar_k)})

  impj2 <-  datlist2mids(jomolist2, progress = F)
  fj20 <- with(impj2, lme4::lmer(Y_ijk ~ Xijktilde20 + Xjkbartilde20 +  Zktilde_20 +  Wk +  (1|
       level3/level2), REML = F))
  est_jm <- t(summary(pool(fj20))$estimate)
  se_jm  <- t(summary(pool(fj20))[,2])
  sum_jm <- c(estimate = est_jm, se = se_jm)
  return(sum_jm)}

jomo2 <- lapply(jomolist2, jomo_x2)
```

```
jomo_sc1 <- data.frame(do.call(rbind, jomo2))
colnames(jomo_sc1)[1:5] <- paste0("est_", names)
colnames(jomo_sc1)[6:10] <- paste0("se_", names)

jomo_bias_sc1 <- (jomo_sc1[1:5] - true_coef)/ true_coef
colnames(jomo_bias_sc1) <- paste0("relbias_", names)
jomo_mse_sc1 <-  (true_coef - jomo_sc1[1:5])^2
colnames(jomo_mse_sc1) <- paste0("mse_", names)

jomo_lowerint <- jomo_sc1[1:5] + qnorm(alpha[1])*jomo_sc1[6:10]
colnames(jomo_lowerint) <- paste0("lower_", names)
jomo_upperint <- jomo_sc1[1:5] + qnorm(alpha[2])*jomo_sc1[6:10]
colnames(jomo_upperint) <- paste0("upper_", names)
jomo_MCse <- jomo_sc1[6:10]/sqrt(nsims)
colnames(jomo_MCse) <- paste0("MCse_", names)

jomo_sc1_comb <- cbind(jomo_sc1, jomo_bias_sc1, jomo_mse_sc1, jomo_lowerint, jomo_upperint,
    jomo_MCse)

write.csv(jomo_sc1_comb, file= "sims_XZ_mcar_jomo_comb.csv", row.names = F)
```

Simulation Code for MCAR in Level 1, 2 and 3 - MICE

```
rm(list=ls(all=TRUE))
pckgs <- c("sn" ,"lme4" , "devtools" ,"miceadds", "extraDistr", "plyr","lmerTest","mice", "
    mitml", "jomo", "truncnorm")

inst = lapply(pckgs, library, character.only = TRUE) # load them
i = 4 # level1 - individuals
j = 20 # level2 - household
k = 100 # level3 - district
## Population intercept
g_000    <- 2
## Population slope
g_100    <- 2.5  # Xijktilde
g_200    <- 2.5 #Xjkbartilde
g_010    <- 2 #Zktilde
g_020    <- 2.5 #zbark
g_001    <- 3 #wk
#### alpha ####
alpha = c(0.025, 0.975)
```

```
true_coef <- c(g_000, g_100, g_200, g_010, g_001)
names <- c("X.Intercept.","Xijktilde","Xjkbartilde","Zktilde", "Wk")
nsims = 1000
simlist = replicate(n = nsims,
                    expr = data.frame(level3 = rep(seq(1:k), each = i*j),
                                      level2 = rep(seq_len(k*j), each = i),
                                      level1 = rep(seq(1:i), each = j*k),
                                      eijk   = rnorm(n = i*j*k, mean = 0, sd = 1),
                                      r0jk   = rep(rnorm(n = k*j, mean = 0, sd = 1), each = i),
                                      u00k   = rep(rnorm(n = k, mean = 0, sd = 1), each = i*j),
                                      X_ijk  = rsn(n = i*j*k, xi= 70, omega = 20, alpha= 10) ,
                                      Z      = rtruncnorm(n = i*j*k, a = 59, b = 396, mean =
                                           114.8, sd = 14),
                                      Zjk    = rep(rtpois(j*k, lambda = 3, a = 2, b = 25),
                                           each = i),
                                      Z1     = rep(rtpois(j*k, lambda = 5.8, a = 1, b = 41),
                                           each = i),
                                      Wk     = rep(rtpois(k, lambda = 3, a = 2, b = 10),  each
                                           = i*j)),
                    simplify = FALSE)
simlist <- lapply(simlist, function(x) within(x, {
  X2 <- (0.6)*Z + sqrt(1-0.6)*Z1
  Z2 <- (0.87)*Zjk + sqrt(1- 0.87)*Z1
  Xbar_jk <- rep(aggregate(X_ijk, list(level2), mean)[,2], each = i)
  Xbar_k <- rep(aggregate(X_ijk, list(level3), mean)[,2], each = i*j)
  Zbar_k <- rep(aggregate(Zjk, list(level3), mean)[,2], each = i*j)
  Xijktilde <- X_ijk - Xbar_jk
  Xjkbartilde <- Xbar_jk - Xbar_k
  Zktilde    <- Zjk - Zbar_k
  Y_ijk <- g_000 + g_100*Xijktilde + g_200*Xjkbartilde +  g_010*Zktilde + g_001*Wk + u00k +
       r0jk + eijk

  #### MCAR in L1 ###########
  Ys <- (Y_ijk - mean(Y_ijk))/sd(Y_ijk)
  X2s <- (X2 - mean(X2))/sd(X2)
  # 20 % missing MCAR
  mcar20x  <- 1- rbinom(i*j*k ,1, 0.20)
  Xijk_20 <- ifelse(mcar20x <1 , NA, X_ijk)
  # 50 % missing MCAR
  mcar50x <- 1- rbinom(i*j*k, 1, 0.50)
  Xijk_50 <- ifelse(mcar50x <1 , NA, X_ijk)

  #### MCAR in L2 ###########
  Yjk <- rep(aggregate(Y_ijk, list(level2), mean)[,2], each = i)
```

```
  Y_sdjk <- rep(aggregate(Y_ijk, list(level2), sd)[,2], each = i)
  Y_s <- (Yjk - mean(Yjk))/sd(Yjk)
  Z2s <- (Z2 - mean(Z2))/sd(Z2)
  mcar20z <- rep(1- rbinom(j*k, 1, 0.20), each = i)
  Zjk_20 <- ifelse(mcar20z <1 , NA, Zjk)
  mcar50z <- rep(1- rbinom(j*k, 1, 0.50), each = i)
  Zjk_50 <- ifelse(mcar50z <1 , NA, Zjk)

  #### MCAR in L3 ###########
  Yk <- rep(aggregate(Y_ijk, list(level3), mean)[,2], each = i*j)
  Y_sdk <- rep(aggregate(Y_ijk, list(level3), sd)[,2], each = i*j)
  Y_sk <- (Yk - mean(Yk))/sd(Yk)
  mcar20w <- rep(1- rbinom(k, 1, 0.20), each = i*j)
  Wk_20 <- ifelse(mcar20w <1 , NA, Wk)
  mcar50w <- rep(1- rbinom(k, 1, 0.50), each = i*j)
  Wk_50 <- ifelse(mcar50w <1 , NA, Wk)
  ############ calculate differences for complete case analysis #############
  Xbarjk_20  <- rep(aggregate(Xijk_20, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_20   <- rep(aggregate(Xijk_20, list(level3), mean, na.rm = T)[,2], each = i*j)
  Xbarjk_50  <- rep(aggregate(Xijk_50, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_50   <- rep(aggregate(Xijk_50, list(level3), mean, na.rm = T)[,2], each = i*j)
  X2barjk    <- rep(aggregate(X2, list(level2), mean)[,2],each = i)
  X2bark     <- rep(aggregate(X2, list(level3), mean)[,2], each = i*j)
  Zbark_20   <- rep(aggregate(Zjk_20, list(level3), mean, na.rm = T)[,2],each = i*j)
  Zbark_50   <- rep(aggregate(Zjk_50, list(level3), mean, na.rm = T)[,2],each = i*j)
  Z2bark     <- rep(aggregate(Z2, list(level3), mean)[,2], each = i*j)

  Xijktilde20 <- Xijk_20 - Xbarjk_20
  Xjkbartilde20 <- Xijk_20 - Xbark_20
  Zktilde_20    <- Zjk_20  - Zbark_20

  Xijktilde50 <- Xijk_50 - Xbarjk_50
  Xjkbartilde50 <- Xijk_50 - Xbark_50
  Zktilde_50    <- Zjk_50  - Zbark_50}))

data <- data.frame(simlist[[1]])
sapply(data, function(x) sum(is.na((x))/nrow(data)))

micelist2 <- simlist
micelist2 <- lapply(micelist2, function(x) { x[c("eijk" , "r0jk","u00k", "X_ijk", "Z" , "Zjk" ,
    "Z1" , "Ys","X2s", "mar20x","mar50x" ,"Xijk_20" ,"Y_sdjk",    "Y_s" , "Z2s", "mar20z","
    mar50z" ,  "Zjk_20","Y_sdk","Y_sk","mar20w","mar50w", "Wk_50",    "Xbarjk_20", "Xbark_20"
    , "Xbarjk_50", "Xbark_50", "Zbark_20"  ,"Zbark_50"  ,"Xijktilde20" , "Xjkbartilde20" , "
    Zktilde_20"  ,  "Xijktilde50" , "Xjkbartilde50" , "Zktilde_50")] <- NULL; x })
```

```
imp_x2 <- function(mice_sim){
  variable_levels <- character(ncol(mice_sim))
  names(variable_levels) <- colnames(mice_sim)
  variable_levels[ c("Zjk_50", "Zktilde", "Xbar_jk", "Xjkbartilde", "Yjk", "Z2", "X2barjk") ]
       <- "level2"
  variable_levels[ c("Wk_20", "Yk", "Xbar_k", "Zbar_k", "X2bark", "Z2bark") ] <- "level3"

  predmat <- make.predictorMatrix(mice_sim)
  predmat[,] <- 0
  predmat[, c("Wk_20", "level3")] <- -2

  predmat["Xijk_50", c("Y_ijk", "X2", "X2barjk", "Zbar_k", "Yjk" )] <-c(3, 3, rep(1,3))
  predmat["Zjk_50", c("Z2", "Yjk","Xbar_jk", "Xbar_k", "Z2bark", "Yk")]<-  c(3,3,3,1,1,1)
  predmat["Wk_20", c("Yk", "Xbar_k", "Zbar_k")]<-  c(3,3,3)

  method <- make.method(mice_sim)
  method[c("Xijk_50", "Zjk_50")] <- "ml.lmer"
  method["Wk_20"] <- "2l.lmer"

  cluster <- list()
  cluster[["Xijk_50"]]<- c("level2", "level3")
  cluster[c("Zjk_50")] <- "level3"

  visit <- c( "Y_ijk", "X2" , "Xbar_k", "Xijk_50", "Xbar_jk", "Xijktilde",
              "Yjk", "Xjkbartilde", "Xbar_jk", "Xbar_k",  "Zjk_50", "Zbar_k", "Zktilde", "Yk","
                   Wk_20")

  imp_x2 <- mice(mice_sim, method=method,levels_id=cluster, variables_levels=variable_levels,
                 predictorMatrix=predmat, visitSequence = visit, seed = 1234)

  impList2 <- miceadds::mids2datlist(imp_x2)
  impList2 <- within(impList2, {
    Xbar_jk <- clusterMeans(Xijk_50, level2)
    Xijktilde50 <- (Xijk_50 - Xbar_jk)
    Xbar_k <- clusterMeans(Xijk_50, level3)
    Xjkbartilde50 <- (Xbar_jk - Xbar_k)
    Zbar_k <- clusterMeans(Zjk_50, level3)
    Zktilde_50   <- (Zjk_50 - Zbar_k)})
  imp50_2 <- datlist2mids(impList2, progress = F)

  lm_mice <- with(imp50_2, lme4::lmer(Y_ijk ~ Xijktilde50 + Xjkbartilde50 +  Zktilde_50 + Wk +
       (1|level3/level2)))
  est_mice <- t(summary(pool(lm_mice))$estimate)
```

```
  se_mice  <- t(summary(pool(lm_mice))[,2])
  sum_mice2  <- c(estimate = est_mice, se = se_mice)
  return(sum_mice2)}

imp2 <- lapply(micelist2, imp_x2)

mice_sc2 <- data.frame(do.call(rbind, imp2))
colnames(mice_sc2)[1:5] <- paste0("est_", names)
colnames(mice_sc2)[6:10] <- paste0("se_", names)
mice_bias_sc2 <- (mice_sc2[1:5] - true_coef)/ true_coef
colnames(mice_bias_sc2) <- paste0("bias_",names)
mice_mse_sc2 <- (true_coef - mice_sc2[1:5])^2
colnames(mice_mse_sc2) <- paste0("mse_",names)

mice_lowerint <- mice_sc2[1:5] + qnorm(alpha[1])*mice_sc2[6:10]
colnames(mice_lowerint) <- paste0("lower_", names)
mice_upperint <- mice_sc2[1:5] + qnorm(alpha[2])*mice_sc2[6:10]
colnames(mice_upperint) <- paste0("upper_", names)
mice_MCse <- mice_sc2[6:10]/sqrt(nsims)
colnames(mice_MCse) <- paste0("MCse_", names)

mice_sc2_comb <- cbind(mice_sc2, mice_bias_sc2, mice_mse_sc2, mice_lowerint, mice_upperint,
    mice_MCse)
write.csv(mice_sc2_comb, file= "sims_XZW_mcar_mice_comb.csv", row.names = F)
```

Simulation Code for MCAR in Level 1, 2 and 3 - JoMo

```
rm(list=ls(all=TRUE))
pckgs <- c("sn" ,"lme4" , "devtools" ,"miceadds", "extraDistr", "plyr","lmerTest","mice", "
    mitml", "jomo", "truncnorm")

inst = lapply(pckgs, library, character.only = TRUE) # load them

i = 4 # level1 - individuals
j = 20 # level2 - household
k = 100 # level3 - district
## Population intercept
g_000    <- 2
## Population slope
g_100    <- 2.5  # Xijktilde
g_200    <- 2.5 #Xjkbartilde
g_010    <- 2 #Zktilde
```

```
g_020    <- 2.5 #zbark
g_001    <- 3 #wk
#### alpha ####
alpha = c(0.025, 0.975)
true_coef <- c(g_000, g_100, g_200, g_010, g_001)
names <- c("X.Intercept.","Xijktilde","Xjkbartilde","Zktilde", "Wk")
nsims = 1
simlist = replicate(n = nsims,
                    expr = data.frame(level3 = rep(seq(1:k), each = i*j),
                                      level2 = rep(seq_len(k*j), each = i),
                                      level1 = rep(seq(1:i), each = j*k),
                                      eijk   = rnorm(n = i*j*k, mean = 0, sd = 1),
                                      r0jk   = rep(rnorm(n = k*j, mean = 0, sd = 1), each = i),
                                      u00k   = rep(rnorm(n = k, mean = 0, sd = 1), each = i*j),
                                      X_ijk  = rsn(n = i*j*k, xi= 70, omega = 20, alpha= 10) ,
                                      Z      = rtruncnorm(n = i*j*k, a = 59, b = 396, mean =
                                           114.8, sd = 14),
                                      Zjk    = rep(rtpois(j*k, lambda = 3, a = 2, b = 25),
                                           each = i),
                                      Z1     = rep(rtpois(j*k, lambda = 5.8, a = 1, b = 41),
                                           each = i),
                                      Wk     = rep(rtpois(k, lambda = 3, a = 2, b = 10),  each
                                           = i*j)),
                    simplify = FALSE)

simlist <- lapply(simlist, function(x) within(x, {
  X2 <- (0.6)*Z + sqrt(1-0.6)*Z1
  Z2 <- (0.87)*Zjk + sqrt(1- 0.87)*Z1
  Xbar_jk <- rep(aggregate(X_ijk, list(level2), mean)[,2], each = i)
  Xbar_k <- rep(aggregate(X_ijk, list(level3), mean)[,2], each = i*j)
  Zbar_k <- rep(aggregate(Zjk, list(level3), mean)[,2], each = i*j)
  Xijktilde <- X_ijk - Xbar_jk
  Xjkbartilde <- Xbar_jk - Xbar_k
  Zktilde    <- Zjk - Zbar_k
  Y_ijk <- g_000 + g_100*Xijktilde + g_200*Xjkbartilde +  g_010*Zktilde + g_001*Wk + u00k +
       r0jk + eijk

  #### MCAR in L1 ###########
  Ys <- (Y_ijk - mean(Y_ijk))/sd(Y_ijk)
  X2s <- (X2 - mean(X2))/sd(X2)
  # 20 % missing MCAR
  mcar20x  <- 1- rbinom(i*j*k ,1, 0.20)
  Xijk_20 <- ifelse(mcar20x <1 , NA, X_ijk)
  # 50 % missing MCAR
```

```
  mcar50x <- 1- rbinom(i*j*k, 1, 0.50)
  Xijk_50 <- ifelse(mcar50x <1 , NA, X_ijk)

  #### MCAR in L2 ###########
  Yjk <- rep(aggregate(Y_ijk, list(level2), mean)[,2], each = i)
  Y_sdjk <- rep(aggregate(Y_ijk, list(level2), sd)[,2], each = i)
  Y_s <- (Yjk - mean(Yjk))/sd(Yjk)
  Z2s <- (Z2 - mean(Z2))/sd(Z2)
  mcar20z <- rep(1- rbinom(j*k, 1, 0.20), each = i)
  Zjk_20 <- ifelse(mcar20z <1 , NA, Zjk)
  mcar50z <- rep(1- rbinom(j*k, 1, 0.50), each = i)
  Zjk_50 <- ifelse(mcar50z <1 , NA, Zjk)

  #### MCAR in L3 ###########
  Yk <- rep(aggregate(Y_ijk, list(level3), mean)[,2], each = i*j)
  Y_sdk <- rep(aggregate(Y_ijk, list(level3), sd)[,2], each = i*j)
  Y_sk <- (Yk - mean(Yk))/sd(Yk)
  mcar20w <- rep(1- rbinom(k, 1, 0.20), each = i*j)
  Wk_20 <- ifelse(mcar20w <1 , NA, Wk)
  mcar50w <- rep(1- rbinom(k, 1, 0.50), each = i*j)
  Wk_50 <- ifelse(mcar50w <1 , NA, Wk)
  ############ calculate differences for complete case analysis #############
  Xbarjk_20  <- rep(aggregate(Xijk_20, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_20   <- rep(aggregate(Xijk_20, list(level3), mean, na.rm = T)[,2], each = i*j)
  Xbarjk_50  <- rep(aggregate(Xijk_50, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_50   <- rep(aggregate(Xijk_50, list(level3), mean, na.rm = T)[,2], each = i*j)
  X2barjk    <- rep(aggregate(X2, list(level2), mean)[,2],each = i)
  X2bark     <- rep(aggregate(X2, list(level3), mean)[,2], each = i*j)
  Zbark_20   <- rep(aggregate(Zjk_20, list(level3), mean, na.rm = T)[,2],each = i*j)
  Zbark_50   <- rep(aggregate(Zjk_50, list(level3), mean, na.rm = T)[,2],each = i*j)
  Z2bark     <- rep(aggregate(Z2, list(level3), mean)[,2], each = i*j)

  Xijktilde20 <- Xijk_20 - Xbarjk_20
  Xjkbartilde20 <- Xijk_20 - Xbark_20
  Zktilde_20    <- Zjk_20  - Zbark_20

  Xijktilde50 <- Xijk_50 - Xbarjk_50
  Xjkbartilde50 <- Xijk_50 - Xbark_50
  Zktilde_50    <- Zjk_50  - Zbark_50}))

data <- data.frame(simlist[[1]])
sapply(data, function(x) sum(is.na((x))/nrow(data)))
jomolist2 <- simlist
jomolist2 <- lapply(jomolist2, function(x) { x[c("eijk" , "r0jk","u00k", "X_ijk", "Z" , "Zjk" ,
```

```
    "Z1" , "Ys","X2s", "mar20x","mar50x" ,"Xijk_20" ,"Y_sdjk","Y_s" , "Z2s", "mar20z","mar50z"
     ,  "Zjk_20","Y_sdk","Y_sk","mar20w","mar50w", "Wk_50",    "Xbarjk_20", "Xbark_20" , "
    Xbarjk_50", "Xbark_50", "Zbark_20"  ,"Zbark_50"  ,   "Xijktilde20" , "Xjkbartilde20" , "
    Zktilde_20")] <- NULL; x })

jomo_x2 <- function(jomo_sim){
  Y = jomo_sim[, c("Xijk_50", "Xijktilde50", "Zjk_50", "Wk_20")]
  clus = jomo_sim[, c("level2")]
  X = jomo_sim[, c("Y_ijk", "X2", "Xjkbartilde", "Zktilde",  "Zbar_k" , "level2", "level3", "Z2
      ")]
  impj <-jomo1ran(Y=Y, X = X, clus = clus , nburn=20, nbetween=20, nimp=10)
  jomo_mids2 <- miceadds::jomo2mids(impj)
  jomolist2 <- miceadds::mids2datlist(jomo_mids2)

  jomolist2 <- within(jomolist2, {
    Xbar_jk <- clusterMeans(Xijk_50, level2)
    Xijktilde50 <- (Xijk_50 - Xbar_jk)
    Xbar_k <- clusterMeans(Xijk_50, level3)
    Xjkbartilde50 <- (Xbar_jk - Xbar_k)
    Zbar_k <- clusterMeans(Zjk_50, level3)
    Zktilde_50 <- (Zjk_50 - Zbar_k)})

  impj2 <-  datlist2mids(jomolist2, progress = F)
  fj20 <- with(impj2, lme4::lmer(Y_ijk ~ Xijktilde50 + Xjkbartilde50 +  Zktilde_50 +  Wk_20 +
      (1|level3/level2), REML = F))
  est_jm <- t(summary(pool(fj20))$estimate)
  se_jm  <- t(summary(pool(fj20))[,2])
  sum_jm <- c(estimate = est_jm, se = se_jm)
  return(sum_jm)}
jomo2 <- lapply(jomolist2, jomo_x2)

jomo_sc1 <- data.frame(do.call(rbind, jomo2))
colnames(jomo_sc1)[1:5] <- paste0("est_", names)
colnames(jomo_sc1)[6:10] <- paste0("se_", names)

jomo_bias_sc1 <- (jomo_sc1[1:5] - true_coef)/ true_coef
colnames(jomo_bias_sc1) <- paste0("relbias_", names)
jomo_mse_sc1 <-  (true_coef - jomo_sc1[1:5])^2
colnames(jomo_mse_sc1) <- paste0("mse_", names)

jomo_lowerint <- jomo_sc1[1:5] + qnorm(alpha[1])*jomo_sc1[6:10]
colnames(jomo_lowerint) <- paste0("lower_", names)
jomo_upperint <- jomo_sc1[1:5] + qnorm(alpha[2])*jomo_sc1[6:10]
colnames(jomo_upperint) <- paste0("upper_", names)
```

```
jomo_MCse <- jomo_sc1[6:10]/sqrt(nsims)
colnames(jomo_MCse) <- paste0("MCse_", names)

jomo_sc1_comb <- cbind(jomo_sc1, jomo_bias_sc1, jomo_mse_sc1, jomo_lowerint, jomo_upperint,
     jomo_MCse)

write.csv(jomo_sc1_comb, file= "sims_XZW_mcar_jomo_comb.csv", row.names = F)
```

Chapter 8

Appendix B

Simulation Code for MAR in Level 1, 2 and 3 - MICE

```
i = 4 # level1 - individuals
j = 20 # level2 - household
k = 100 # level3 - district
## Population intercept
g_000    <- 2
## Population slope
g_100    <- 2.5  # Xijktilde
g_200    <- 2.5 #Xjkbartilde
g_010    <- 2 #Zjktilde
g_020    <- 2.5 #zbark
g_001    <- 3 #wk
#### alpha ####
alpha = c(0.025, 0.975)
true_coef <- c(g_000, g_100, g_200, g_010, g_001)
names <- c("X.Intercept.","Xijktilde","Xjkbartilde","Zjktilde", "Wk")
nsims = 1000
simlist = replicate(n = nsims,
                    expr = data.frame(level3 = rep(seq(1:k), each = i*j),
                                      level2 = rep(seq_len(k*j), each = i),
                                      level1 = rep(seq(1:i), each = j*k),
                                      eijk   = rnorm(n = i*j*k, mean = 0, sd = 1),
                                      r0jk   = rep(rnorm(n = k*j, mean = 0, sd = 1), each = i),
                                      u00k   = rep(rnorm(n = k, mean = 0, sd = 1), each = i*j),
                                      X_ijk  = rsn(n = i*j*k, xi= 70, omega = 20, alpha= 10) ,
                                      Z      = rnorm(n = i*j*k, mean = 114.8, sd = 14),
                                      Zjk    = rep(rtpois(j*k, lambda = 3, a = 2, b = 25),
                                           each = i),
                                      Z1     = rep(rtpois(j*k, lambda = 5.8, a = 1, b = 41),
                                           each = i),
                                      Wk     = rep(rtpois(k, lambda = 3, a = 2, b = 10),  each
```

```
                                        = i*j)),
                    simplify = FALSE)
simlist <- lapply(simlist, function(x) within(x, {
  X2 <- (0.6)*Z + sqrt(1-0.6)*Z1
  Z2 <- (0.87)*Zjk + sqrt(1- 0.87)*Z1
  Xbar_jk <- rep(aggregate(X_ijk, list(level2), mean)[,2], each = i)
  Xbar_k <- rep(aggregate(X_ijk, list(level3), mean)[,2], each = i*j)
  Zbar_k <- rep(aggregate(Zjk, list(level3), mean)[,2], each = i*j)
  Xijktilde <- X_ijk - Xbar_jk
  Xjkbartilde <- Xbar_jk - Xbar_k
  Zjktilde    <- Zjk - Zbar_k
  Y_ijk <- g_000 + g_100*Xijktilde + g_200*Xjkbartilde +  g_010*Zjktilde + g_001*Wk + u00k +
       r0jk + eijk

  #### MAR in L1 ###########
  Ys <- (Y_ijk - mean(Y_ijk))/sd(Y_ijk)
  X2s <- (X2 - mean(X2))/sd(X2)
  mar50x <- round(exp(X2s +(0.85)*Ys)/ 1 + exp(X2s + (0.85)*Ys),0)
  Xijk_50 <- ifelse(mar50x > quantile(mar50x)[3][[1]] , NA, X_ijk)

  #### MAR in L2 ###########
  Yjk <- rep(aggregate(Y_ijk, list(level2), mean)[,2], each = i)
  Y_sdjk <- rep(aggregate(Y_ijk, list(level2), sd)[,2], each = i)
  Y_s <- (Yjk - mean(Yjk))/sd(Yjk)
  Z2s <- (Z2 - mean(Z2))/sd(Z2)
  mar50z <- round(exp(Z2s+(-0.85)*Ys)/ 1 + exp(Z2s + (-0.85)*Ys),0)
  Zjk_50 <- ifelse(mar50z > quantile(mar50z)[3][[1]] , NA, Zjk)

  #### MAR in L3 ###########
  Yk <- rep(aggregate(Y_ijk, list(level3), mean)[,2], each = i*j)
  Y_sdk <- rep(aggregate(Y_ijk, list(level3), sd)[,2], each = i*j)
  Y_sk <- (Yk - mean(Yk))/sd(Yk)
  mar20w <-round(exp(2+(-0.85)*Y_sk)/ 1 + exp(2 + (-0.85)*Y_sk),0)
  Wk_20 <- ifelse(mar20w > quantile(mar20w)[4][[1]], NA, Wk)

  ############ calculate differences for complete case analysis #############
  Xbarjk_50  <- rep(aggregate(Xijk_50, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_50   <- rep(aggregate(Xijk_50, list(level3), mean, na.rm = T)[,2], each = i*j)
  X2barjk    <- rep(aggregate(X2, list(level2), mean)[,2],each = i)
  X2bark     <- rep(aggregate(X2, list(level3), mean)[,2], each = i*j)
  Zbark_50   <- rep(aggregate(Zjk_50, list(level3), mean, na.rm = T)[,2],each = i*j)
  Z2bark     <- rep(aggregate(Z2, list(level3), mean)[,2], each = i*j)
  Xijktilde50 <- Xijk_50 - Xbarjk_50
  Xjkbartilde50 <- Xijk_50 - Xbark_50
```

```
  Zjktilde_50    <- Zjk_50  - Zbark_50}))

################## MICE ##########################
micelist2 <- simlist
micelist2 <- lapply(micelist2, function(x) { x[c("eijk" , "r0jk","u00k", "X_ijk", "Z" , "Zjk" ,
     "Z1" , "Ys","X2s", "mar20x","mar50x" ,"Xijk_20" ,"Y_sdjk",  "Y_s" , "Z2s", "mar20z","
     mar50z" ,  "Zjk_20","Y_sdk","Y_sk","mar20w","mar50w", "Wk_50",    "Xbarjk_20", "Xbark_20"
     , "Xbarjk_50", "Xbark_50", "Zbark_20"  ,"Zbark_50"  , "Xijktilde20" , "Xjkbartilde20" , "
     Zjktilde_20"  ,  "Xijktilde50" , "Xjkbartilde50" , "Zjktilde_50")] <- NULL; x })

imp_x2 <- function(mice_sim){
  variable_levels <- character(ncol(mice_sim))
  names(variable_levels) <- colnames(mice_sim)
  variable_levels[ c("Zjk_50", "Zjktilde", "Xbar_jk", "Xjkbartilde", "Yjk", "Z2", "X2barjk") ]
      <- "level2"
  variable_levels[ c("Wk_20", "Yk", "Xbar_k", "Zbar_k", "X2bark", "Z2bark") ] <- "level3"

  predmat <- make.predictorMatrix(mice_sim)
  predmat[,] <- 0
  predmat[, c("Wk_20", "level3")] <- -2
  predmat["Xijk_50", c("Y_ijk", "X2", "X2barjk", "Zbar_k", "Yjk" )] <-c(3, 3, rep(1,3))
  predmat["Zjk_50", c("Z2", "Yjk","Xbar_jk", "Xbar_k", "Z2bark", "Yk")]<-  c(3,3,3,1,1,1)
  predmat["Wk_20", c("Yk", "Xbar_k", "Zbar_k")]<-  c(3,3,3)

  method <- make.method(mice_sim)
  method[c("Xijk_50", "Zjk_50")] <- "ml.lmer"
  method["Wk_20"] <- "2l.lmer"

  cluster <- list()
  cluster[["Xijk_50"]]<- c("level2", "level3")
  cluster[c("Zjk_50")] <- "level3"

 visit <- c( "Y_ijk", "X2" , "Xbar_k", "Xijk_50", "Xbar_jk", "Xijktilde",   "Yjk", "Xjkbartilde
      ", "Xbar_jk", "Xbar_k",  "Zjk_50", "Zbar_k", "Zjktilde", "Yk","Wk_20")

#  model <- list(Xijk_20 = "continuous", Zjk_20 = "pmm")

imp_x2 <- mice(mice_sim, method=method,levels_id=cluster, variables_levels=variable_levels,
                 predictorMatrix=predmat, visitSequence = visit, seed = 1234)

  impList2 <- miceadds::mids2datlist(imp_x2)
  impList2 <- within(impList2, {
    Xbar_jk <- clusterMeans(Xijk_50, level2)
    Xijktilde50 <- (Xijk_50 - Xbar_jk)
```

```
    Xbar_k <- clusterMeans(Xijk_50, level3)
    Xjkbartilde50 <- (Xbar_jk - Xbar_k)
    Zbar_k <- clusterMeans(Zjk_50, level3)
    Zjktilde_50   <- (Zjk_50 - Zbar_k)})
  imp50_2 <- datlist2mids(impList2, progress = F)

  lm_mice <- with(imp50_2, lme4::lmer(Y_ijk ~ Xijktilde50 + Xjkbartilde50 +  Zjktilde_50 + Wk +
        (1|level3/level2)))
  est_mice <- t(summary(pool(lm_mice))$estimate)
  se_mice  <- t(summary(pool(lm_mice))[,2])
  sum_mice2  <- c(estimate = est_mice, se = se_mice)
  return(sum_mice2)}

imp2 <- lapply(micelist2, imp_x2)

mice_sc2 <- data.frame(do.call(rbind, imp2))
colnames(mice_sc2)[1:5] <- paste0("est_", names)
colnames(mice_sc2)[6:10] <- paste0("se_", names)
mice_bias_sc2 <- (mice_sc2[1:5] - true_coef)/ true_coef
colnames(mice_bias_sc2) <- paste0("bias_",names)
mice_mse_sc2 <- (true_coef - mice_sc2[1:5])^2
colnames(mice_mse_sc2) <- paste0("mse_",names)

mice_lowerint <- mice_sc2[1:5] + qnorm(alpha[1])*mice_sc2[6:10]
colnames(mice_lowerint) <- paste0("lower_", names)
mice_upperint <- mice_sc2[1:5] + qnorm(alpha[2])*mice_sc2[6:10]
colnames(mice_upperint) <- paste0("upper_", names)
mice_MCse <- mice_sc2[6:10]/sqrt(nsims)
colnames(mice_MCse) <- paste0("MCse_", names)

mice_sc2_comb <- cbind(mice_sc2, mice_bias_sc2, mice_mse_sc2, mice_lowerint, mice_upperint,
    mice_MCse)
write.csv(mice_sc2_comb, file= "sims_XZW_mar_mice1_comb.csv", row.names = F)
```

Simulation Code for MAR in Level 1, 2 and 3 - JoMo

```
i = 4 # level1 - individuals
j = 20 # level2 - household
k = 100 # level3 - district
## Population intercept
g_000    <- 2
## Population slope
```

```
g_100    <- 2.5  # Xijktilde
g_200    <- 2.5 #Xjkbartilde
g_010    <- 2 #Zjktilde
g_020    <- 2.5 #zbark
g_001    <- 3 #wk
#### alpha ####
alpha = c(0.025, 0.975)
true_coef <- c(g_000, g_100, g_200, g_010, g_001)
names <- c("X_Intercept","Xijktilde","Xjkbartilde","Zjktilde", "Wk")
nsims = 1000
simlist = replicate(n = nsims,
                    expr = data.frame(level3 = rep(seq(1:k), each = i*j),
                                      level2 = rep(seq_len(k*j), each = i),
                                      level1 = rep(seq(1:i), each = j*k),
                                      eijk   = rnorm(n = i*j*k, mean = 0, sd = 1),
                                      r0jk   = rep(rnorm(n = k*j, mean = 0, sd = 1) , each = i)
                                          ,
                                      u00k   = rep(rnorm(n = k, mean = 0, sd = 1), each = i*j),
                                      X_ijk  = rsn(n = i*j*k, xi= 70, omega = 20, alpha= 10) ,
                                      Z      = rnorm(n = i*j*k, mean = 114.8, sd = 14),
                                      Zjk    = rep(rtpois(j*k, lambda = 3, a = 2, b = 25),
                                          each = i),
                                      Z1     = rep(rtpois(j*k, lambda = 5.8, a = 1, b = 41),
                                          each = i),
                                      Wk     = rep(rtpois(k, lambda = 3, a = 2, b = 10),  each
                                          = i*j)),
                    simplify = FALSE)

simlist <- lapply(simlist, function(x) within(x, {
  X2 <- (0.6)*Z + sqrt(1-0.6)*Z1
  Z2 <- (0.87)*Zjk + sqrt(1- 0.87)*Z1
  Xbar_jk <- rep(aggregate(X_ijk, list(level2), mean)[,2], each = i)
  Xbar_k <- rep(aggregate(X_ijk, list(level3), mean)[,2], each = i*j)
  Zbar_k <- rep(aggregate(Zjk, list(level3), mean)[,2], each = i*j)
  Xijktilde <- X_ijk - Xbar_jk
  Xjkbartilde <- Xbar_jk - Xbar_k
  Zjktilde    <- Zjk - Zbar_k
  Y_ijk <- g_000 + g_100*Xijktilde + g_200*Xjkbartilde +  g_010*Zjktilde + g_001*Wk + u00k +
       r0jk + eijk

  #### MAR in L1 ###########
  Ys <- (Y_ijk - mean(Y_ijk))/sd(Y_ijk)
  X2s <- (X2 - mean(X2))/sd(X2)
```

```
  mar50x <- round(exp(X2s +(0.85)*Ys)/ 1 + exp(X2s + (0.85)*Ys),0)
  Xijk_50 <- ifelse(mar50x > quantile(mar50x)[3][[1]] , NA, X_ijk)

  #### MAR in L2 ###########
  Yjk <- rep(aggregate(Y_ijk, list(level2), mean)[,2], each = i)
  Y_sdjk <- rep(aggregate(Y_ijk, list(level2), sd)[,2], each = i)
  Y_s <- (Yjk - mean(Yjk))/sd(Yjk)
  Z2s <- (Z2 - mean(Z2))/sd(Z2)
  mar50z <- round(exp(Z2s+(-0.85)*Ys)/ 1 + exp(Z2s + (-0.85)*Ys),0)
  Zjk_50 <- ifelse(mar50z > quantile(mar50z)[3][[1]] , NA, Zjk)

  #### MAR in L3 ###########
  Yk <- rep(aggregate(Y_ijk, list(level3), mean)[,2], each = i*j)
  Y_sdk <- rep(aggregate(Y_ijk, list(level3), sd)[,2], each = i*j)
  Y_sk <- (Yk - mean(Yk))/sd(Yk)
  mar20w <-round(exp(2+(-0.85)*Y_sk)/ 1 + exp(2 + (-0.85)*Y_sk),0)
  Wk_20 <- ifelse(mar20w > quantile(mar20w)[4][[1]], NA, Wk)

  ############ calculate differences for complete case analysis #############
  Xbarjk_50  <- rep(aggregate(Xijk_50, list(level2), mean, na.rm = T)[,2],each = i)
  Xbark_50   <- rep(aggregate(Xijk_50, list(level3), mean, na.rm = T)[,2], each = i*j)
  X2barjk    <- rep(aggregate(X2, list(level2), mean)[,2],each = i)
  X2bark     <- rep(aggregate(X2, list(level3), mean)[,2], each = i*j)

  Zbark_50   <- rep(aggregate(Zjk_50, list(level3), mean, na.rm = T)[,2],each = i*j)
  Z2bark     <- rep(aggregate(Z2, list(level3), mean)[,2], each = i*j)

  Xijktilde50 <- Xijk_50 - Xbarjk_50
  Xjkbartilde50 <- Xijk_50 - Xbark_50
  Zjktilde_50    <- Zjk_50  - Zbark_50}))

jomolist2 <- simlist
jomolist2 <- lapply(jomolist2, function(x) { x[c("eijk" , "r0jk","u00k", "X_ijk", "Z" , "Zjk" ,
    "Z1" , "Ys","X2s", "mar20x","mar50x" ,"Xijk_20" ,"Y_sdjk", "Y_s" , "Z2s", "mar20z","mar50z
    " , "Zjk_20","Y_sdk","Y_sk","mar20w","mar50w", "Wk_50",     "Xbarjk_20", "Xbark_20" , "
    Xbarjk_50", "Xbark_50", "Zbark_20"  ,"Zbark_50"  ,  "Xijktilde20" , "Xjkbartilde20" , "
    Zjktilde_20")] <- NULL; x })

jomo_x2 <- function(jomo_sim){
  Y = jomo_sim[, c("Xijk_50", "Xijktilde50", "Zjk_50", "Wk_20")]
  clus = jomo_sim[, c("level2")]
  X = jomo_sim[, c("Y_ijk", "X2", "Xjkbartilde", "Zjktilde",  "Zbar_k" , "level2", "level3", "
      Z2")]
  impj <-jomo1ran(Y=Y, X = X, clus = clus , nburn=20, nbetween=20, nimp=10)
```

```
  jomo_mids2 <- miceadds::jomo2mids(impj)
  jomolist2 <- miceadds::mids2datlist(jomo_mids2)

  jomolist2 <- within(jomolist2, {
    Xbar_jk <- clusterMeans(Xijk_50, level2)
    Xijktilde50 <- (Xijk_50 - Xbar_jk)
    Xbar_k <- clusterMeans(Xijk_50, level3)
    Xjkbartilde50 <- (Xbar_jk - Xbar_k)
    Zbar_k <- clusterMeans(Zjk_50, level3)
    Zjktilde_50 <- (Zjk_50 - Zbar_k)})

  impj2 <-  datlist2mids(jomolist2, progress = F)
  fj20 <- with(impj2, lme4::lmer(Y_ijk ~ Xijktilde50 + Xjkbartilde50 +  Zjktilde_50 +  Wk_20 +
        (1|level3/level2), REML = F))
  est_jm <- t(summary(pool(fj20))$estimate)
  se_jm  <- t(summary(pool(fj20))[,2])
  sum_jm <- c(estimate = est_jm, se = se_jm)
  return(sum_jm)}

jomo2 <- lapply(jomolist2, jomo_x2)

jomo_sc1 <- data.frame(do.call(rbind, jomo2))
colnames(jomo_sc1)[1:5] <- paste0("est_", names)
colnames(jomo_sc1)[6:10] <- paste0("se_", names)

jomo_bias_sc1 <- (jomo_sc1[1:5] - true_coef)/ true_coef
colnames(jomo_bias_sc1) <- paste0("relbias_", names)
jomo_mse_sc1 <-  (true_coef - jomo_sc1[1:5])^2
colnames(jomo_mse_sc1) <- paste0("mse_", names)

jomo_lowerint <- jomo_sc1[1:5] + qnorm(alpha[1])*jomo_sc1[6:10]
colnames(jomo_lowerint) <- paste0("lower_", names)
jomo_upperint <- jomo_sc1[1:5] + qnorm(alpha[2])*jomo_sc1[6:10]
colnames(jomo_upperint) <- paste0("upper_", names)
jomo_MCse <- jomo_sc1[6:10]/sqrt(nsims)
colnames(jomo_MCse) <- paste0("MCse_", names)

jomo_sc1_comb <- cbind(jomo_sc1, jomo_bias_sc1, jomo_mse_sc1, jomo_lowerint, jomo_upperint,
     jomo_MCse)
write.csv(jomo_sc1_comb, file= "sims_XZW_mar_jomo_comb.csv", row.names = F)
```