Causal Mediation Analysis with A Partially Missing Mediator: Exploring the Effect of Portable Air Purifier Use During Pregnancy on Infant Birth Weight

by

Xi S. Kang

B.A. (Health Sciences), Simon Fraser University, 2019
B.B.A. (Marketing), Thompson Rivers University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for The Degree of
Master of Science

in the
Master of Science Program
Faculty of Health Sciences

© Xi S. Kang 2021
SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Xi S. Kang**
Degree:	**Master of Science (Health Sciences)**
Title:	**Causal Mediation Analysis with A Partially Missing Mediator: Exploring the Effect of Portable Air Purifier Use During Pregnancy on Infant Birth Weight**
Committee:	**Chair: Jeremy Snyder** Professor, Health Sciences **Lawrence McCandless** Supervisor Professor, Health Sciences **Ryan Allen** Committee Member Professor, Health Sciences **Jennifer Hutcheon** Committee Member Associate Professor, Obstetrics & Gynaecology University of British Columbia **Hui Xie** Examiner Professor, Health Sciences

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Mediation analysis examines the exposure-outcome association that acts through an intermediate variable. However, mediation analysis becomes challenging when data have missing values. Although methods exist to deal with missing data and mediation analysis independently, few studies have examined how to combine the approaches, specifically, how to pool the mediation analysis results across a series of imputed datasets and compute confidence intervals for target parameters. We propose a new technique that combines multiple imputation with maximum likelihood estimation. Using computer simulations, we compare the performance of our proposed approach with a traditional bootstrap approach. Our method performs well and is more computationally efficient than other resampling methods. We apply the new method to randomized trial data on the role of cadmium exposure in mediating the effects of an environmental health intervention on birth weight.

Keywords: Multiple Imputation; Mediation Analysis; MICE; Maximum Likelihood; Biomarkers; Maternal Health; Birth Weight; UGAAR

Acknowledgements

I would like to start by thanking everyone on my supervisory committee. This work would not have been finished without their guidance and support. Thank you, Lawrence McCandless and Ryan Allen, for your constant encouragement and guidance, as well as Jennifer Hutcheon, for your valuable and professional suggestions. Biggest thanks to Lawrence for spending much time reading through each draft and providing me with inspiring advice. I would also like to thank all the collaborators at the UGAAR study. I am grateful to be part of this program and analyze this data.

Thanks to my best friend Felicia Leong for always having her faith in me. She is the best friend one could ever ask for. Thanks, Yahui, for the past decade. Special thanks to my precious and beloved dog, Arsene, for the endless love and trust.

This work was supported by funding provided by the Ulaanbaatar Gestational and Air Pollution Research study.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AM:	Intervention-mediator
AY:	Intervention-outcome
BMI:	Body mass index
BCa:	Bias-corrected and accelerated
BCCI:	Bias-corrected confidence interval
BOOT(MI):	Multiple imputation nested within bootstrap
CC:	Complete case
CDE:	Controlled direct effect
CGM:	Conditional Gaussian Model
CI:	Confidence interval
DAG:	Directed acyclic graph
DE:	Direct effect
FCS:	Fully conditional specification
FMI:	Fraction of missing information
HEPA:	High-efficiency particulate air
IE:	Indirect effect
ITT:	Intention-to-treat
JM:	Joint modelling
MAR:	Missing at random
MCAR:	Missing completely at random

MI:	Multiple imputation
MI(BOOT):	Bootstrap nested within multiple imputation
MICE:	Multivariate Imputation by Chained Equations
MICE(ML):	MICE combined with maximum likelihood
MLE:	Maximum likelihood estimates
MNAR:	Missing not at random
MC:	Monte Carlo
MY:	Mediator-outcome
NDE:	Natural direct effect
NIE:	Natural indirect effect
PM:	Proportion mediated
RCT:	Randomized clinical trial
TE:	Total effect
UGAAR:	Ulaanbaatar Gestation and Air Pollution Research

Chapter 1 Introduction

1.1 Background

Mediation analysis is a tool that helps researchers examine the underlying association between an exposure and an outcome through an intermediate variable, which is called a mediator (Baron & Kenny, 1986). For example, if a mediator (M) is the variable in a causal pathway between the exposure (A) and the outcome (Y), then A causes M and M causes Y.

Mediation analysis, which has gained popularity in environmental epidemiology, serves to quantify the contribution of environmental factors to the associations between exposures and health outcomes (VanderWeele, 2016). More specifically, motivations for evaluating mediation effects include strengthening evidence of the main effect hypothesis, understanding the pathway through which exposure causes disease, and evaluating and improving interventions (Hafeman & Schwartz, 2009).

In environmental epidemiology, biomarkers are often used to measure the exposure to and/or risk from environmental toxicants (Travis, 1993). Examples of biomarkers include chemical (e.g., lead) concentrations in human tissues such as blood or urine. Biomarkers have the advantage that they give a precise estimate of exposure from multiple pathways. However, a difficulty with using biomarkers in environmental epidemiology is that biomarkers are often not completely recorded for all study participants. Little attention and research has addressed the causes of missing data (Mfutso-Bengo, Masiye, Molyneux, Ndebele, & Chilungo, 2008). Nonetheless, at the data analysis stage, it is common for epidemiologists to simply discard data records with missing information and use complete-case (CC) analysis, but this can cause biased estimation and loss of statistical power (Graham, 2009).

An important example of missing biomarkers comes from the Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study. The UGAAR study is an ongoing randomized air purifier intervention birth cohort study in Ulaanbaatar, Mongolia, focused on the impacts of portable indoor high-efficiency particulate air (HEPA) filters during pregnancy on fetal growth and early childhood development. In previous work, Barn et al. (2018) reported that HEPA purifier use during pregnancy was associated with a 14% (95% CI: 4, 23%) reduction in maternal blood cadmium

(Barn, Gombojav, Ochir, Boldbaatar, et al., 2018) and an 85 g (95% CI: 3, 167 g) greater mean birth weight among babies born at term (Barn, Gombojav, Ochir, Laagan, et al., 2018). In a more recent study, the investigators found that a doubling of blood cadmium was associated with a 95 g (95% CI: 34, 155 g) and 91 g (95% CI: 32, 150 g) reduction in average birth weight among all births and term births, respectively, in adjusted models (Barn et al., 2019). However, the authors reported that almost 20% of participants had missing values in blood cadmium concentration (Barn et al., 2019).

1.2 Challenges and objectives

1.2.1 Challenge 1: Missing data and multiple imputation

In their study of the association between gestational cadmium exposure and birth weight, Barn et al. (2019) included only data from 374 participants with a blood cadmium measurement from a total of 463 live births. In addition, a CC analysis eliminated an additional 64 participants with missing data in other adjustment variables, such as parity and household income, leaving only 310 participants entered the analytic model. Thus, in total, more than 30% of participants were excluded from the analysis. The 310 complete cases constitute a subset of the original sample. Therefore, in the UGAAR analysis of cadmium exposure and birth weight, the use of CC analysis may have reduced statistical power and the precision of effect estimates, and its results may not be comparable to and representative of the original cohort.

Furthermore, participants in the intervention group were more likely to consent to provide biomarkers (odds ratio: 1.88, 95% CI: 1.07 - 3.30) (Barn et al., 2019). There were 240 live births in the intervention group, of which 203 (84.6%) mothers provided blood samples. In contrast, 174 out of 223 (78%) participants with live births in the control group provided cadmium measurements. This illustrates that the missing data does not occur in a way that is completely at random, and instead it depends on participant characteristics, such as assignment to intervention or control.

While the UGAAR study provides one example of missing data, it is a widespread problem in epidemiologic research. In an assessment that examined 278 molecular epidemiology studies, almost all (95%) either had missing data on one or more variables or used the availability of data as the condition for final cohort selection (Desai, Kubo, Esserman, & Terry, 2011). In some studies, biomarker data were missing due not only to limits of detection (LOD) but also incomplete collection (M. Lee et al., 2018).

To evaluate the effect of missing data on the estimated association between cadmium and birth weight reported by Barn et al. (2019), the first objective of this MSc thesis was to use the MICE algorithm to generate the missing values for cadmium and any remaining variables from the UGAAR cohort. We hypothesize that using MICE to handle the missing values and re-analyzing the full UGAAR data with imputed values will confirm the association between maternal blood cadmium concentrations and birth weight by making the regression coefficient estimates more precise, correcting for bias from the CC analysis and lowering the standard errors.

1.2.2 Challenge 2: Assessing the mediating role of blood cadmium

Previously, Barn and co-workers found HEPA purifiers reduced average blood cadmium concentration (Barn, Gombojav, Ochir, Laagan, et al., 2018), HEPA purifier use increased birth weight (Barn, Gombojav, Ochir, Boldbaatar, et al., 2018), and blood cadmium was associated with decreased birth weight (Barn et al., 2019). Putting all these findings together, the second objective of this MSc thesis was to investigate the potential mediator role of cadmium in the relationship between HEPA purifier use and birth weight, for example, to estimate the overall total effect of HEPA purifier use on birth weight, the direct effect of HEPA purifier use going straight on birth weight, the indirect effect that goes through blood cadmium, and the proportion of this mediated effect. The definition of these effects will be described later in this thesis.

1.2.3 Challenge 3: Incorporating multiple imputation into a mediation analysis

Furthermore, a crucial challenge is combining mediation analysis with MI. Specifically, it is unknown how to conduct a mediation analysis across a series of data imputations (for example, how to calculate standard errors for the indirect effects that incorporate imputation uncertainty). There is little guidance in the literature about how to combine inferences from mediation analysis generated from multiple imputed datasets. Additionally, the function in software designed to process mediation analysis on multiple datasets where missing values have been imputed is not completely developed yet (Tingley et al., 2019). Therefore, the third objective of this MSc thesis was to develop a general approach to mediation analysis with missing values in the dataset that is computationally efficient and can be applied to different settings.

Chapter 2 Literature review

2.1 Missing data mechanisms

The mechanisms of missing data can be classified into three types depending on the probability of a value being missing: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little & Rubin, 2019; Rubin, 1976, 2009). MCAR occurs when the probability of missingness is unrelated to either the value of other observed or unobserved variables, or the values of the variable itself. MAR occurs if the probability of missingness is related to the observed values of other variables in the dataset but not to the variable itself. This is the mechanism that is most commonly assumed. MNAR occurs when the probability of missingness is related to the values of the missing variable itself. For example, people with very high income usually do not want to answer questions to disclose their income.

The missing data mechanism has an effect on the choice of method(s) for handling missingness. CC analysis (also called listwise deletion) is the default in many statistical packages and the most commonly used analysis method analysis. Using CC analysis, observations with missing values are simply deleted from the dataset. Generally, CC analysis gives unbiased estimates of means and variances if the data are MCAR, which means that the complete data are a random sample drawn from the study population. The disadvantage of CC analysis lies in the reduction in sample size and enlarged standard errors (Bouhlila & Sellaouti, 2013). In reality, the MCAR case rarely occurs.

In contrast, if the data are MAR or MNAR, then CC analysis can produce severely biased results. In this case, the complete cases may not be comparable to and representative of the original cohort. Thus, CC analysis may lead to invalid association with biased parameter estimates and large standard errors. In particular, the bias increases with the proportion of missingness and the differences between observed and missing values (Stef van Buuren, 2018). Some studies showed that CC analysis may result in a risk of overestimated benefits and underestimated harm (Sterne et al., 2009).

2.2 Multiple imputation (MI), and the MICE method

Multiple imputation (MI) has become a popular method to handle missing data in epidemiological studies, especially in dealing with more than one variable with missingness (Rubin, 1996). The general idea of MI is to create K complete datasets by performing K independent imputations and then to replace the missing values with plausible data values (Johnson & Young, 2011). Two general approaches for generating imputed multivariate data are joint modelling (JM) and fully conditional specification (FCS). The mechanism of JM will not be discussed here as it is not the focus of this paper, and it has seen less widespread use in epidemiology. FCS specifies many multivariate imputation models for each incomplete variable by a set of conditional densities on a variable-by-variable basis (Stef van Buuren & Groothuis-Oudshoorn, 2011b).

Imputation analysis is superior to CC analysis because imputation not only utilizes the full dataset but also produces smaller standard errors and narrower confidence intervals (CI) (Bouhlila & Sellaouti, 2013). It uses more information from the dataset by incorporating the auxiliary variables in the imputation model that are not used in the analytical model (Lall, 2016). While CC analysis requires missing data to be MCAR, MI is based on the MAR assumption and may lead to unbiased results (Graham, 2009). Another merit of MI is that the multiple imputed datasets account for the statistical uncertainty in the imputations by calculating the between-imputation variances (Azur, Stuart, Frangakis, & Leaf, 2011).

MI using the basic idea of FCS has been proposed in various names, such as stochastic relaxation (Kennickell, 1991), variable-by-variable imputation (J. J. Brand, 1999), regression switching (S. van Buuren, Boshuizen, & Knook, 1999), and the chained equations (Van Buuren & Oudshoorn, 2000). Chained equations, initially released as an S-PLUS library in 2000, has become one popular approach to MI, namely Multivariate Imputation by Chained Equations (MICE) (Van Buuren & Oudshoorn, 2000).

MICE is particularly popular because it is very flexible to use. Many of the initially developed MI methods are based on the assumption of a large joint model for all the variables, for example, a joint normal distribution (Azur et al., 2011). In the MICE procedure, each variable is modeled conditional on the other variables, which means that each variable can be modeled according to its own distribution. Therefore, MICE can be applied in many different data types (e.g., binary variables using logistic regression and continuous variables using linear regression). The algorithm of MICE in detail will be introduced in the following section.

2.3 The MICE algorithm

During MICE, each imputed dataset is generated after several iterations of the imputation algorithm. The MICE algorithm starts from iteration zero where missing values are filled by a random draw from the observed values of each variable (White, Royston, & Wood, 2011).

There are several steps in iteration one. First, variable 1 with missing values is set back to missing and is used as the dependent variable in a regression model. This model is called the imputation model, which uses information from all other variables in the model. Details of the imputation model will be discussed in section 2.3.1. In this step the values for the other variables with missingness are used from iteration zero. Therefore, no variables have missing values except the variable 1. Here, the predicted (imputed) values of variable 1 are obtained.

In the next step, the same procedure is repeated for the second variable with missing values, which is set back to missing and is regressed on the other variables in the imputation model, including the variable 1. Variable 1 uses the imputed values in the last step.

The third step to impute the third variable with missingness is similar to the previous two steps. The missing values in variable 1 and variable 2 are replaced by the predicted values imputed in the previous two steps. Iteration one finishes when all variables with missing values have been cycled through. The number of steps in iteration one is equal to the number of variables with missing values. Because each variable is imputed with its own imputation model, MICE has the ability to handle different types of variables (e.g. continuous, binary, categorial, etc.) (White et al., 2011).

Having completed iteration one, the MICE algorithm moves on to iteration two. Iteration two starts with variable 1 with similar procedures as iteration one, but the rest of the variables with missingness use the imputed values from the last step of iteration one. Same steps are taken to impute the rest of variables with missing values.

The iteration process is repeated several times. The proper number of iterations is explained in section 2.3.2. The imputed values in the final iteration, with the observed data in the original dataset, are used as the first imputed dataset.

This entire process is repeated for the next imputed dataset. After K repetitions, we obtain K imputed datasets. The proper number of imputations (K) will be discussed in section 2.3.3. The quantity of scientific interest, for example, the regression coefficient, can be estimated on each of the imputed datasets. The K estimates and their variances are then pooled into one estimate (Stef van Buuren & Groothuis-Oudshoorn, 2011b).

2.3.1 Imputation models

Another gain provided by MI over CC analysis is that MI can take advantage of the information that is not part of the planned analysis. We use imputation models to fill the missing values with random imputations. Then we analyze the full dataset with imputed values using the analytic model.

The imputation model may be different than the analytic model, but should include all of the variables in the analytic model, and perhaps some more variables (White et al., 2011). Stuart suggests stepwise selection to choose the most predictive variables (Stuart, Azur, Frangakis, & Leaf, 2009).

More commonly, researchers choose to include auxiliary variables in the multiple imputation. They are variables in the dataset that are not included in the analytic model (i.e. not predictive of the outcome variable), but are either correlated with a missing variable or believed to be associated with missingness (von Hippel & Lynch, 2013). For example, researchers may include participants' weight or height to impute missing body mass index (BMI). The inclusion of auxiliary variables improves the efficiency of MI estimates obtained from the analytical model because it enables precise estimation of the missing information (Enders, 2010). Adding auxiliary variables in the imputation models may also make the MAR assumption more plausible and improve the imputations.

The variables in the imputation model are not required to be in a specific order because the software imputes variables in order from the one with the least missingness to the one with the most missingness, by default (Bouhlila & Sellaouti, 2013).

2.3.2 Number of iterations

The number of iterations in RStudio is set to be 5, by default, but researchers can use a higher number as needed. Generally, a low number of 5 to 20 iterations seems to be enough to reach convergence and excessive iterations can slow the computational speed (S. van Buuren et al., 1999). Convergence monitoring can be done by plotting the variance between imputations and variance within imputations. van Buuren believes no more than 5 iterations per imputation usually produce unbiased estimates and appropriate coverage (Stef van Buuren, 2018).

2.3.3 Number of multiple imputations

At first, Rubin suggests that 3 to 5 imputations are enough for multiple imputation (Rubin, 1996). Rubin's suggestion was based on the efficiency of point estimates and did not address imputation variability. Researchers nowadays are interested in not only efficiency for point estimates, but also standard error estimates, confidence intervals, and p-values.

From a Monte Carlo simulation (Graham, Olchowski, & Gilreath, 2007), we can see that as the number of imputations goes to infinity, the precision of the pooled parameter becomes as large as possible and regression coefficients are essentially unbiased. The fewer imputations, the more loss of power and precision.

Many researchers have different recommendations on the grounds of retaining the power for testing an association of interest. Graham suggests at least 20 imputations to restrict the loss of power (Graham et al., 2007). Bodner considers the fraction of missing information (FMI) and concludes the required number of imputed datasets depending on different FMI (Bodner, 2008). White, Royston and Wood (2011) quote a rule of thumb from Von Hippel (P. T. von Hippel, 2009) that the number of imputed datasets should be at least equal to the percentage of missing cases. For example, if 10% of the participants have missing values, then at least 10 imputed datasets should be generated. van Buuren commented that theoretically, higher number of imputations mean more precise results, but the substantive conclusions will not change as a result of raising the number beyond 5 (Stef van Buuren, 2018).

2.4 Mediation analysis

To illustrate the main ideas of mediation analysis, Figure 1 is a simple diagram that shows how the total effect of A on Y is separated into a direct effect relating A straight to Y and a mediated effect where A has an indirect effect on Y through M. The upper half of this figure represents the total effect (through ϕ) that A can have on Y. The lower half of this figure shows the pathways that, A leads to M (through β) and M lead to Y (through θ_2). It also shows the relationship between A and Y controlling for M, representing the direct effect, which is denoted by θ_1.

2.4.1 Mediation analysis approaches

There are two common analytical approaches to conduct the mediation analysis: statistical and causal (H. Lee, Herbert, & McAuley, 2019). Statistical mediation analysis commonly uses the product/difference methods to test indirect effects. It uses regression models to estimate the exposure-mediator and mediator-outcome association. The indirect effect is denoted by either the product of β and θ_2, or the difference of ϕ (i.e., total effect) and θ_1 (i.e., direct effect). For continuous outcome and mediator, these indirect effects estimated by the difference and product methods should be the same (i.e. $\beta\theta_2 = \varphi - \theta_1$) (VanderWeele, 2016). The causal mediation analysis, commonly called the Causal Inference approach, is not model specific and it defines direct and indirect effect from counterfactuals or a potential outcomes framework (Pearl, 2001).

The product method has two main limitations. It only works in the special cases where there are linear relationships between the exposure, mediator, and outcome and when there is no exposure-mediator interaction (MacKinnon, 2012). The causal inference approach was developed to address these limitations. It is applicable to nonlinear models with both discrete and continuous variables. It also allows exposure-mediator interaction (Pearl, 2001).

The statistical (product/difference methods) and causal (causal inference approach) mediation approaches produce equal estimates when the association between the exposure, mediator, and outcome are linear and when there is no exposure-mediator interaction. When mediator or outcome variables are binary variables or when interactions are present, statistical mediation can produce biased estimates and, causal mediation is preferred (VanderWeele & VanderWeele, 2015).

It is important to test the significance of the mediation effect to see whether the mediated effect is significantly different from zero. The Sobel test was often used to test the significance of the product of β and θ_2 (Sobel, 1982). The test is given by dividing the estimate of the product by an approximate estimate of the standard error of the product derived via the Delta method, and then comparing the ratio with a standard normal distribution. The confidence intervals are also calculated based on the critical values from the standard normal distribution (Sobel, 1982). However, the power of the Sobel test is questionable due to the non-normality of the distribution of the product. The product of two normally distributed variables is usually not normally distributed (Springer & Thompson, 1966), and therefore the 95% confidence interval should not be symmetric. There is evidence that this traditional method to test indirect effect has imbalanced confidence limits and results in low statistical power and type I error rates (Mackinnon, Lockwood, & Williams, 2004).

However, this issue can be overcome with bootstrap resampling. This method is a nonparametric resampling procedure with replacement and does not rely on the distribution assumption on the indirect effect (Mackinnon et al., 2004). With the empirical distribution of $\beta\theta_2$, a confidence interval, a p-value, or a standard error can be determined using percentiles of the distribution. If the confidence interval does not contain zero, the researcher can conclude that the indirect effect is different from zero. The bias-corrected and accelerated (BCa) intervals by this nonparametric bootstrapping serves the optimal result as it offers the most statistical power to detect a mediating effect while keeping the Type I error rate within the robustness interval (Mackinnon et al., 2004).

To use the product method to test the mediation effects, there are some specific assumptions that need to be met: 1) No measurement error in variables (Hoyle & Kenny, 1999), 2) The causal relations of exposure to mediator to outcome are correctly specified (McDonald, 1997), 3) No omitted variables (McDonald, 1997), and 4) No interaction of exposure and mediator (Judd & Kenny, 1981). In addition, a mediation analysis should be free of intervention-outcome (AY), intervention-mediator (AM), and mediator-outcome (MY) confounding effects to make valid causal inferences (VanderWeele & VanderWeele, 2015). This is called the no unmeasured confounding assumption.

In a randomized controlled trial, participants are randomly assigned to the exposure groups (intervention or control group). This means that the intervention is statistically independent of the outcome, mediator, and all covariates. Therefore, the intervention-outcome (AY) and intervention-mediator (AM) effects can be assumed to be unconfounded (H. Lee et al., 2019). However, the mediator-outcome (MY) effect may be confounded, even in a randomized controlled trial, because participants are not randomly assigned to the mediator level. For example, smoking is a confounding variable for the association between cadmium levels and birth weight because smoking causes higher cadmium levels (Järup & Åkesson, 2009), and additionally, smoking also causes lower birth weight babies (Bernstein et al., 2005). Thus, it is important to adjust for smoking in order to estimate the causal effect of cadmium on birth weight. To avoid the potential bias, we need to control for all available confounders of the mediator-outcome effect. To assess the potential bias caused by unmeasured/unavailable confounders, sensitivity analysis should be employed to examine the robustness of mediated effect.

2.5 A gap in the biostatistics literature: Combining MI with mediation analysis.

An important concern is how to combine MI with mediation analysis. Specifically, the challenge is to combine the mediation analysis results across a series of imputed datasets and test the corresponding statistical significance using bootstrap.

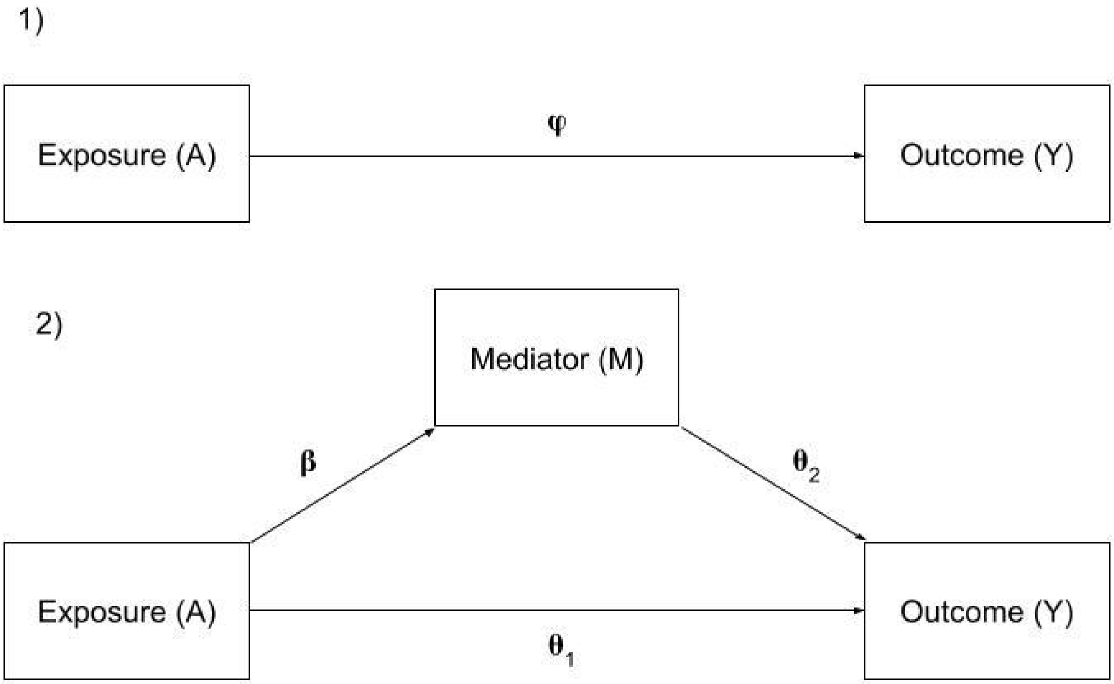
In addition, the software function to process mediation analysis on multiple imputed data sets is not completely developed; there are only two published papers exploring this issue through simulation studies (Wu & Jia, 2013; Z. Zhang, Wang, & Tong, 2015). Although they shared the same idea, the authors of these two articles proposed opposing strategies. Wang et al. (2015) propose to make MI nested within bootstrap samples, whereas Wu and Jia (2013) use bootstrap nested within MI.

Wang et al. (2015) started with drawing a bootstrap sample of N persons randomly with replacement from the original data set (sample size = N) and this resampling data set would include missing data. Then they performed MI (with K imputed data sets) on the bootstrap sample and obtained the K point estimates (e.g., θ) of the mediation effect. No standard error was calculated for each point estimate. The K point estimates were then pooled into one estimate by taking their mean. These steps were repeated for a total of B bootstrap samples to get B mediation effect coefficients. The B coefficients constituted an empirical distribution, and the confidence intervals of the mediation effect were constructed using percentiles of this distribution. One of the limitations of this technique is that it is computationally impractical because it requires MI possibly hundreds of times (once for each bootstrap sample). This procedure is denoted by BOOT(MI). A further limitation is that Wang et al. (2015) only considered multiple imputation of multivariate normally missing data. In theory, one could instead use MICE to do the imputations, and this would resolve this limitation.

In contrast, Wu and Jia (2013) proposed an opposite procedure: First use MI to get K imputed data sets and then apply the bootstrap (B bootstrap samples) to each imputed dataset. This procedure is more practical and is denoted by MI(BOOT). The K times B estimates were combined into a single large frequency distribution to approximate the mediation effects and the bias-corrected confidence interval (BCCI) can then be calculated. Their rationale was grounded on that their procedure is more computationally efficient than Wang et al.'s. MI is much more computationally costly than bootstrap. Using bootstrap nested within MI, they process, say 1 run of MI and 1000 runs of bootstrap, instead of 1000 runs of MI and 1 run of bootstrap by MI nested within bootstrap. This will be more computationally efficient and will take significantly less computing time when dealing with big data with thousands of variables. Wu and Jia (2013) also compared these two procedures using a real dataset and both showed comparable point estimates and bias-corrected confidence intervals.

Apart from these two papers (Wu & Jia, 2013; Z. Zhang et al., 2015), there has been no work examining the details of how exactly to combine MI with mediation analysis. Moreover, we stress that this topic has never been discussed in the context of missing biomarkers of exposure for environmental epidemiology. Consequently, it is an important research direction that needs further study.

Figure 1: A simple path diagram of mediation model



Chapter 3 Manuscript

3.1 Introduction

Mediation analysis techniques examine the association between an exposure and outcome that acts through an intermediate variable (i.e., the mediator). The total effect of the exposure on the outcome can be decomposed as the indirect effect via the mediator and the direct effect relating the exposure straight to the outcome (Baron & Kenny, 1986). Mediation analysis is widely used in many research fields, including the social sciences (Hoven & Siegrist, 2013; Rucker, Preacher, Tormala, & Petty, 2011), epidemiology (Bind et al., 2014; Nasiri, Moodie, & Abenhaim, 2020), and behaviour research (Bonnert et al., 2018; Morgan, Mackinnon, & Jorm, 2013). Mediation analysis is particularly useful in environmental epidemiology studies that use biomarkers of environmental chemical exposures. These measurements take the form of chemical concentrations measured in blood, urine or other biospecimens. Biomarkers have the advantage of providing objective estimates of chemical exposures from multiple sources. Thus, they help explain the complex pathways among environmental exposures, biological processes and health outcomes (Hu, Zhuang, Bernardo, & McCandless, 2018).

However, a limitation of biomarkers in mediation analysis is that they are often not completely recorded for all study participants, which leads to missing values in the data. In a recent study that examined 278 published molecular epidemiology studies, a total of 66% had missing data for at least one biomarker (Desai et al., 2011). Of these studies with missing biomarkers, only 12% used some type of missing data methods, and all of these studies used single imputation and/or use of missing data indicators. The remaining 88% studies used complete case (CC) analysis to deal with the missing data (Desai et al., 2011). CC analysis is the default in many statistical packages and the most common practice in epidemiology studies (Bouhlila & Sellaouti, 2013). It can produce severely biased results because the complete cases left in the analysis may not be representative of the original cohort (Graham, 2009). In addition, CC analysis is cost-inefficient because many valuable data are deleted. Single imputation using the predicted mean value is also known to have potentially poor performance because it does not incorporate the uncertainty of the imputation (J. Brand, van Buuren, le Cessie, & van den Hout, 2019).

To deal with missing data, a widely used approach is multiple imputation (MI). MI is automated in many statistical software packages (Royston, 2004; SAS Institute, n.d.; Stata, n.d.). The general idea of MI is to create m complete datasets by performing independent imputations and then to replace the missing values with plausible data values (Johnson & Young, 2011). MI is superior to CC analysis because imputation not only utilizes the full dataset but also frequently produces smaller standard errors and narrower confidence intervals (Bouhlila & Sellaouti, 2013). Furthermore, the multiple imputed datasets account for the statistical uncertainty in the imputations by incorporating the within- and between-imputation variances (Azur et al., 2011).

Mediation analysis presents a host of unique challenges to MI. When taking missing values into account, a critical challenge is deciding how to combine the mediation analysis results across a series of imputed datasets and then compute 95% confidence intervals for target parameters. It has been well documented about how to calculate standard errors and 95% confidence intervals for the indirect effect with complete data (e.g. Sobel test (Sobel, 1982) and bootstrapping (Bollen & Stine, 1990; Shrout & Bolger, 2002)). However, there is little guidance in the literature about how to combine inferences from mediation analysis generated from multiple imputed datasets.

To our knowledge, there are only two published papers that examine mediation analysis in multiple imputation settings, including Zhang et al. (Z. Zhang et al., 2015) and Wu and Jia (Wu & Jia, 2013). Both papers developed similar approaches that rely on bootstrap resampling to combine MI and the mediation analysis (Wu & Jia, 2013; Z. Zhang et al., 2015). Although both studies showed good performance in simulation studies, they have some limitations in the application. After calculating the indirect effect, both studies described the magnitude of the effect sizes simply by categorizing them as small, medium, or large, following Cohen standards (Cohen, 1988). They did not report any quantitative effect size. Furthermore, both approaches are computationally intensive because a large number of bootstrap samples usually takes a long time to process. Crucially, both studies of Wu & Jia, 2013 and Z. Zhang et al., 2015 only considered multiple imputation settings for multivariate normally missing data. Thus, the methods are less well suited to handle the settings where the dataset contains different types of missing data (e.g., continuous, binary, categorical data).

Recently, the *amelidiate* and *mediations* functions in the *mediation* package in RStudio have been developed to conduct mediation analysis among multiple imputed datasets (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). These functions share a similar algorithm with the approach proposed by Wu and Jia (Wu & Jia, 2013). With these functions, they impute the missing data first, then apply the mediation analysis to the imputed datasets with bootstrapped standard errors, and then combine the resulting inferences in order to average over imputations. However, the *amelidiate* function remains under-developed. They do not yield p-values for hypothesis testing, and they do not support models with ordered outcome (Tingley et al., 2019).

Combining mediation analysis results and multiple imputation is an important topic for environmental epidemiology, and new methods are needed. Accordingly, the motivation of this study is to develop a novel approach to mediation analysis with missing values in the dataset that is computationally efficient and can be applied to different missing data settings. To achieve this goal, we use Multivariate Imputation by Chained Equations (MICE), combined with Maximum Likelihood (ML) estimation using Monte Carlo (MC) simulation. This approach has the advantage that it is less computationally intensive than the existing approaches (Tingley et al., 2019; Wu & Jia, 2013; Z. Zhang et al., 2015) that use the bootstrap, and additionally, it is easy to implement and can accommodate different data types of missing data.

In the discussion that follows, we introduce a motivating example from the Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study, its data description and some preliminary results in Section 3.2. Section 3.3 reviews the existing mediation analysis methods, the MICE algorithm, and the details of our newly proposed approach that combines MICE with ML. In Section 3.4, we apply the mediation analysis to UGAAR study, and we examine the effects of portable air purifier use and maternal blood cadmium concentration on infant birth weight. We examine the mediating role of blood cadmium. In Section 3.5, we illustrated the results of a simulation study, comparing the different approaches to mediation analysis with missing data. Finally, we make conclusions, discussion of limitations and future research directions in Section 3.6.

3.2 Motivating example: Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study

We illustrate the challenges of mediation analysis in environmental epidemiology with missing biomarkers using data from the UGAAR study. The UGAAR study is an ongoing randomized air purifier intervention birth cohort study in Ulaanbaatar, Mongolia, focused on the impacts of portable indoor high-efficiency particulate air (HEPA) purifier use during pregnancy on fetal growth and early childhood development. Ulaanbaatar, Mongolia is one of the most polluted cities in the world and it is highly suited to examine the health benefits of HEPA purifiers use (UNICEF, n.d.). In the present study, we investigated the effect of the portable HEPA purifiers use during pregnancy on infant birth weight, and additionally, we examined the mediating role of blood cadmium concentration measured in the blood of study participants.

3.2.1 Data description

The UGAAR study population included pregnant women ≥ 18 years old, and ≤ 18 weeks into a single-gestation pregnancy. The women were non-smokers, living in an apartment, planning to give birth in an Ulaanbaatar maternity hospital, and not using a portable air purifier in the home at the time of enrolment. Full details regarding the cohort and the data can be found in the study by Barn et al.(Barn, Gombojav, Ochir, Laagan, et al., 2018).

Figure 2 shows the trial profile, including 540 participants who were randomly allocated to HEPA intervention or control group. The intervention group received one or two HEPA purifiers, based on apartment size, to use from enrolment to the end of pregnancy. The control group received no air purifier. After excluding 28 women who withdrew consent or moved out of study area, there were 512 women who were followed until the end of pregnancy. Among the 512 women, there were a total of 47 pregnancy losses, 2 chromosomal abnormalities, and 463 participants with live births. 6 women were further excluded from the study due to self-report on smoking in late pregnancy. In all of the analyses that follow, except Section 4.4.2, we limited our analyses to the 457 live births.

Table 1 shows basic descriptive statistics for the n = 457 live births. There were a total of 423 term births (i.e., gestational age ≥ 37 weeks) and 34 preterm births. Participants in control and intervention groups had similar demographic characteristics and lifestyles, including age at enrolment, household income, and pre-pregnancy BMI (Table 1).

An important focus of the study was examining biomarkers of cadmium exposure. As illustrated in Table 1, out of the 457 live births, a total of 84 (~20%) participants did not provide blood samples, and they had missing values in blood cadmium concentrations. Furthermore, some other important variables were often missing, including income, pre-pregnancy BMI, and parity, leading to less than 70% of the cohort having complete data.

3.2.2 Preliminary analysis results and scientific question

In previous analyses of the UGAAR data, Barn et al. (Barn, Gombojav, Ochir, Boldbaatar, et al., 2018; Barn, Gombojav, Ochir, Laagan, et al., 2018) reported that air purifier use during pregnancy was associated with a 14% (95% CI: 4, 23%) reduction in maternal blood cadmium and, additionally, an 85 g (95% CI: 3, 167 g) increase in mean birth weight among babies born at term. More recently, Barn et al. (Barn et al., 2019) found that a doubling of blood cadmium concentration was associated with a 95 g (95% CI: 34, 155 g) and 91 g (95% CI: 32, 150 g) reduction in average birth weight among all births and term births, respectively, in adjusted models. These findings are demonstrated in Table 1, where we observe that the intervention group had a 100 g higher median birth weight, and additionally, a 0.03 μg/L lower median blood cadmium concentration, compared to the control group.

Putting all these findings together, our scientific questions are as follows: 1) Does cadmium mediate the relationship between HEPA purifier use during pregnancy and infant birth weight, and, if so, what proportion of the health benefits of HEPA purifier use can be contributed to cadmium, and 2) Is our proposed MICE(ML) method, described below, a better approach in dealing with missing values in mediation analysis than using CC analysis with bootstrap, in terms of producing more precise estimates, with lower computational cost compared to published methods (Wu & Jia, 2013; Z. Zhang et al., 2015).

3.3 Methodology

3.3.1 Mediation analysis

3.3.1.1 General definitions of mediation effects

In this section, we review two common analytical approaches to mediation analysis: statistical mediation analysis versus causal mediation analysis (H. Lee et al., 2019; Pearl, 2012; Vanderweele, 2015). Statistical mediation analysis uses linear models and it is commonly used for continuous outcomes and mediators with linear regression models and no exposure-mediator interaction (VanderWeele, 2016). In contrast, causal mediation analysis defines direct and indirect effect using the potential outcomes framework. It is often used as the extension of the statistical approach to allow for nonlinearity, exposure-mediator interaction, and categorical data (Pearl, 2001).

Statistical mediation analysis expresses the direct and indirect effects in terms of the regression coefficients associated with the models for the outcome and the mediator. Let A be the exposure, Y the outcome, M the mediator and C the vector of covariates, the two models are regressed as follows:

$$E[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_4' c. \qquad [1]$$

$$E[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c. \qquad [2]$$

If the models are correctly specified, the direct effect (DE) is equal to the parameter θ_1, and it is interpreted as the effect of the exposure on the outcome conditional on mediator level and other covariates controlled at a specific value.(MacKinnon, 2012) The indirect effect (IE) is equal to the quantity $\beta_1\theta_2$, which is the product of the exposure coefficient in Equation [2] and the mediator coefficient in Equation [1]. Therefore, this approach is generally referred to as the product method. The IE is interpreted as the effect of the exposure on the outcome that passes through the mediator, conditional on covariates. The total effect (TE) is composed of the sum of DE (θ_1) and IE ($\beta_1\theta_2$), representing the overall effect of the exposure on the outcome.

To describe the causal mediation analysis approach, originally defined by Robin & Greenland (Robins & Greenland, 1992) and Pearl (Pearl, 2001), let $Y(a,m)$be the potential outcome of Y, possibly counterfactual to the fact, if A was set to a and M was set to m. If A is binary, then $a = 1$ and $a^* = 0$. Let $M(a)$ be the potential outcome of M that we would see if A was set to a.

The controlled direct effect (CDE) comparing the effect of A on Y when A = a with A = a^* and when the mediator is fixed at level m is defined by $Y(a,m) - Y(a^*,m)$. The average CDE(m) for a population, conditional on covariates C = c, is denoted by $\mathbb{E}(Y(a,m) - Y(a^*,m)|c)$.

In contrast, the natural direct effect (NDE) is formally defined by $Y(a, M(a^*)) - Y(a^*, M(a^*))$. It expresses the effect of A on Y by changing the exposure from a^* to a, but setting the mediator level to $M(a^*)$, which is what it naturally would have been if exposure had been a^*. Likely, we define the average NDE for the population conditional on covariates C = c , $\mathbb{E}\big(Y\big(a, M(a^*)\big) - Y\big(a^*, M(a^*)\big)\big|c\big)$.

Correspondingly, the natural indirect effect (NIE) expresses how much the outcome would change on average when the exposure was fixed at level a but the mediator was changed from the level it would have been if exposure had been a^* to the level it would have been if exposure had been a. The average NIE for the population, conditional on covariates C = c, is formally defined as $\mathbb{E}\big(Y\big(a, M(a)\big) - Y\big(a, M(a^*)\big)\big|c\big)$. It is easily to see that the total effect is composed of the sum of the NDE and NIE: $Y(a) - Y(a^*) = Y\big(a, M(a)\big) - Y\big(a^*, M(a^*)\big) = \big[Y\big(a, M(a) - Y\big(a, M(a^*)\big)\big] + \big[Y\big(a, M(a^*)\big) - Y\big(a^*, M(a^*)\big)\big]$.

For valid parameter estimation in causal mediation analysis, several assumptions are required, including:

A1: There is no unmeasured confounder for the exposure-outcome relationship.

A2: There is no unmeasured confounder for the mediator-outcome relationship.

A3: There is no unmeasured confounder for the exposure-mediator relationship.

A4: There is no mediator-outcome confounder is itself affected by the exposure.

A1 and A2 are required to estimate CDE and A3 and A4 are two additional assumptions for estimating NDE and NIE. A1 and A3 would be automatically satisfied when the study is a randomized clinical trial (RCT), i.e., the exposure is randomly assigned. Note that even with an RCT, A2 is not guaranteed to be satisfied.

Applying these definitions into practice, we limit our attention to the case of a continuous outcome (birth weight) and mediator ($\log_2$-transformed cadmium concentration), and binary exposure variable (HEPA purifier use) as in the UGAAR study. When the exposure and the mediator interact in their effect on the outcome, these counterfactual effects can be estimated from the regression models and Equation [1] is replaced by

$$E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4{}' c. \qquad [3]$$

The causal inference approach also applies to the scenarios in which one or both of the outcome and mediator are binary. See VanderWeele (VanderWeele, 2016) for further details.

If the models are correctly specified and the confounding Assumptions A1–A4 hold, the controlled direct effect (CDE), natural direct effect (NDE), and natural indirect effect (NIE) estimates for a change in the exposure from level a to a^* are given as follows:

$$CDE(m) = (\theta_1 + \theta_3 m)(a - a^*)$$

$$NDE = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$$

$$NIE = (\beta_1\theta_2 + \beta_1\theta_3 a)(a - a^*)$$

Note that in the absence of AM interaction (i.e., $\theta_3 = 0$), the CDE and the NDE are equal to the DE (i.e., θ_1) obtained using statistical approach, and the NIE is equal to the IE (i.e., $\beta_1\theta_2$) obtained using statistical approach.

3.3.1.2 Effect size measures

The majority of mediation analysis studies do not report a quantitative effect size for the mediated effect (Klumparendt, Nelson, Barenbrügge, & Ehring, 2019; Murphy, Shevlin, Houston, & Adamson, 2012; J. Zhang, Wedel, & Pieters, 2009). Of those which do report the effect size, the proportion mediated (PM) is currently the most commonly used measure to assess the relative magnitude of the indirect effect (Preacher & Kelley, 2011). It is also the only and the default effect size measure in the *mediation* package in RStudio (Tingley et al., 2019). The PM is calculated by dividing the indirect effect by the total effect:

$$PM = \beta_1\theta_2/(\beta_1\theta_2 + \theta_1).$$

Although the PM has some limitations, such as that its confidence interval is often quite wide, and additionally, that it lacks a clear interpretation when the direct and indirect effects are in opposite direction, it is nonetheless a helpful summary of how important the mediator is in explaining the association of the exposure on the outcome (VanderWeele & VanderWeele, 2015)。

3.3.2 Review of interval estimation for model parameters in mediation analysis

The confidence intervals are calculated based on the critical values from the standard normal distribution. Calculating 95% confidence intervals for model parameters can be challenging in mediation analysis, particularly for estimating the indirect effect because it is a non-linear function of the parameters in Equations [1] and [2] (Mackinnon et al., 2004; Preacher & Hayes, 2004; Sobel, 1982). The most commonly used methods include the Sobel test, bootstrapping, and Monte Carlo method (Bollen & Stine, 1990; Mackinnon et al., 2004; Shrout & Bolger, 2002; Sobel, 1982). The Sobel test obtained via the delta method is a conservative but the most commonly used estimate. The Sobel standard error of the product $\beta_1\theta_2$ is estimated by $\sqrt{\theta_2^2 SE_{\beta_1}^2 + \beta_1^2 SE_{\theta_2}^2}$. However, the power of the Sobel test is questionable because the product of two normally distributed variables is usually not normally distributed, especially when the sample size is small, and therefore the 95% confidence interval should not be symmetric (Preacher & Hayes, 2004). Bootstrapping, as a nonparametric resampling procedure does not rely on the distribution assumption on the indirect effect. From each of the bootstrapped samples, the indirect effect is computed and an empirical sampling distribution of $\beta_1\theta_2$ is derived (Mackinnon et al., 2004; Preacher & Hayes, 2004; Shrout & Bolger, 2002). Monte Carlo method uses the estimates of β_1 and θ_2 and their standard errors and generates random normal variables for β_1 and θ_2 to create an empirical distribution of $\beta_1\theta_2$ (Mackinnon et al., 2004). However, the application of these methods has been limited to studies with complete data. The presence of missing values has put another layer of challenge to handle the nonnormal distribution of $\beta_1\theta_2$.

3.3.3 Multivariate Imputation by Chained Equations

Before describing our proposed method for mediation analysis with missing data, we briefly review MICE for multiple imputation. MICE has become a popular approach to MI because of its flexibility to use. MICE, also known as “fully conditional specification”, imputes each variable with missing values conditional on all the other variables. Unlike other MI approaches which assume a large joint model for all the variables (e.g., a joint normal distribution), MICE models each variable according to its own distribution. Therefore, MICE can handle many different data types (e.g., binary variables using logistic regression and continuous variables using linear regression) (Stef van Buuren & Groothuis-Oudshoorn, 2011a). The implementation of MICE assumes the missing data are missing at random (MAR), which means the probability of missingness is related to the observed values of other variables in the dataset but not to the variable itself (Rubin, 1976).

The algorithm of MICE starts from a random draw from the observed values of each variable to replace the missing values (White et al., 2011). Then, the first variable with missing values to be imputed will be set back to missingness and this variable will be regressed on all the other variables. The missing values are replaced by the predictive values from the posterior predictive distribution. This procedure is cycled through each variable with missingness to finish a single iteration. Such an iteration will be repeated several times (typically 5 to 10) to stabilize the results and to produce a single imputed dataset. The entire process is repeated independently m times, obtaining m imputed datasets, which have identical observed data entries but different imputed values. See Section 6 for discussion of the appropriate number of imputations that should be used with MICE. Within each imputed dataset, the quantity of interest is estimated, and these estimates differ for each imputed dataset because of the different imputed values. The m estimates are pooled into one single estimate using Rubin's Rules (Carpenter, 2013) with commensurate standard error, thereby combining the variation within and between the m imputed datasets.

3.3.4 Combining MICE and Maximum Likelihood Estimation in a mediation analysis

In this section, we propose an approach that combines MICE and Maximum Likelihood estimation in order to obtain point and interval estimates for target parameters in the mediation analysis. We call this method MICE(ML), in contrast to MI(BOOT) and BOOT(MI) methods described by Wu and Jia (2013) and Zhang et al. (2015). Our proposed MICE(ML) can be related to the theory of the multivariate delta method described by Sobel (Sobel, 1982),` that is used for calculating standard errors for estimators (e.g., the indirect effect) that are nonlinear functions of asymptotically normal random variables. To illustrate the main idea, note that, the quantities of interest in the mediation analysis are θ_1 (direct effect), $\beta_1\theta_2$ (indirect effect), $\theta_1 + \beta_1\theta_2$ (total effect) and $\beta_1\theta_2/(\theta_1 + \beta_1\theta_2)$ (the proportion mediated). In finite samples, the resulting estimators will not be normally distributed because they are non-linear combinations of asymptotically normally distributed estimators for θ_1, θ_2, and β_1 (Sobel, 1982).

Rather than using the bootstrap, we can calculate 95% CIs for the four quantities using Monte Carlo simulation by generating from a normal distribution approximation of the maximum likelihood estimates (MLE) $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\beta}_1$ obtained by the regression models from Equation 1 and 2.

The following is the description of the MICE(ML) procedure for mediation analysis with a partially missing mediator. Figure 3 is a visualized presentation of these steps and we used ψ as a general parameter for convenience.

1. MICE step: Apply MICE to the incomplete dataset with missing values to obtain m imputed datasets, imp_i, $i = 1, \dots, m$.
2. For i in $1{:}m$
    a. Analysis step: Using Equations [1] and [2], estimate θ_1, θ_2, and β_1 within the i^{th} imputed dataset and obtain ML estimates $\hat{\theta}_{1l}$, $\hat{\theta}_{2l}$, and $\hat{\beta}_{1l}$, and their standard errors $se(\hat{\theta}_{1l})$, $se(\hat{\theta}_{2l})$, and $se(\hat{\beta}_{1l})$.
3. For j in $1{:}n$
    a. ML simulation step: From a random normal distribution approximation of the MLE for each parameter estimate, simulate n random numbers for each parameter, denoted $\hat{\theta}_{1_{ij}}$, $\hat{\theta}_{2_{ij}}$, and $\hat{\beta}_{1_{ij}}$
4. Combination step: These $m \times n$ MC simulations of $\widehat{\theta_1}$, $\widehat{\theta_2}$, and $\widehat{\beta_1}$ are then used to generate the sampling distributions of $\widehat{\beta_1\theta_2}$ (indirect effect), $\widehat{\theta_1} + \widehat{\beta_1\theta_2}$ (total effect) and $\widehat{\beta_1\theta_2}$ /$(\widehat{\theta_1} + \widehat{\beta_1\theta_2})$ (the proportion mediated). The lower and upper confidence limits for each quantity are obtained from the corresponding percentiles of the empirical distributions.

The combination step is justified because it is a numerical integration of the target quantities over the posterior distribution of the missing data (van de Schoot et al., 2014).

This proposed approach has several advantages. First, it accounts for the statistical uncertainty due to missing data in the imputations. MICE can handle different types of missing data so that it can be applied to many different settings. Second, the uncertainty due to random sampling error is accounted for by ML. Third, ML is much more computationally efficient than other resampling methods, such as bootstrap. This advantage can be further prominent when there are a large number of variables or samples in the dataset. Fourth, the 95% confidence intervals constructed by taking the quantiles of the empirical distributions is easy to understand, implement and interpret.

3.4 Analysis of the UGAAR data: Assessing the mediating role of maternal blood cadmium concentration in the relation between air purifier use and infant birth weight

3.4.1 Methodology

We applied the MICE(ML) method to the UGAAR Study data. To re-iterate, the exposure is HEPA purifier use during pregnancy, the outcome is infant birth weight, and the mediator is blood cadmium concentration among pregnant women, which is partially missing. We used a log-2 transformation to handle the skewness blood cadmium concentration and reduce the influence of outliers (Romano, Enquobahrie, Simpson, Checkoway, & Williams, 2016; VanderWeele, 2016). The base-2 logarithm was used for easy interpretation.

Following the multi-step procedure described in Section 3.3.4, we first applied MICE to the UGAAR data. The primary variable with missingness of interest is the mediator. However, building on the analyses of Barn et al. (Barn et al., 2019; Barn, Gombojav, Ochir, Boldbaatar, et al., 2018; Barn, Gombojav, Ochir, Laagan, et al., 2018), there were 5 other explanatory variables with missingness, including living with a smoker in late pregnancy, pre-pregnancy BMI, parity, income, and coal stove density within 5,000 meters. In total, there were more than 30% of subjects had missing values in at least one variable. We used the *mice* package in RStudio (Stef van Buuren & Groothuis-Oudshoorn, 2011a) to impute each variable in a model conditional on the outcome, exposure, mediator, and all the rest of the covariates. Based on the conclusions by previous researchers (Bodner, 2008; Graham et al., 2007; Stef van Buuren, 2018; White et al., 2011), we used 5 iterations per imputation and create 20 imputed datasets.

Following the multi-step procedure, we fit the analytical models in Equations 1 and 2 to the imputed datasets. We did not include exposure mediator interactions in the model because no significant interaction was detected. When selecting adjustment variables in the analytic models, we used a causal directed acyclic graph (DAG) (see Figure 4) to present the relations between the exposure, mediator, outcome and a set of covariates. The selected covariates were either potential mediator-outcome confounders (e.g., living with a smoker in late pregnancy), or those which may have an impact on the mediator or outcome (e.g., coal stove density and sex of baby). Because the DAG implies the causal relationships between variables, we adjust for the following covariates: Maternal age at birth (<25, 25-29, 30-34, ≥35 years), pre-pregnancy BMI (kg/m^2, continuous), monthly household income (<600,000, 600,000 to <1,000,000, ≥1,000,000 Tugriks), living with a smoker in late pregnancy (no, yes), anemia (no, yes), parity (0, 1, ≥2), coal stove density within 5,000 m (gers/hectare, continuous), and sex of the baby (female, male). Additionally, following Barn et al. (Barn, Gombojav, Ochir, Boldbaatar, et al., 2018), when modeling birth weight we also adjust for gestational age and gestational age squared (weeks, continuous).

We conducted the mediation analysis in several different ways. Firstly, we compared MICE(ML) with the CC analysis, because the CC method is widely used in the literature. Secondly, in order to isolate the effect of the intervention on fetal growth, we presented an analysis of all births, and additionally, an analysis that was restricted to term births, which are defined as a birth with gestational age that was greater than 37 weeks. With the CC analysis, the 95% CIs for the IE, TE, and PM were calculated with 10,000 iterations of bootstrap resampling. With MICE(ML), 10,000 simulations are generated from the MLEs computed from each of the 20 imputed datasets. 95% CIs for the IE, TE, and PM were calculated from the empirical distribution of the 200,000 MLs simulations.

3.4.2 Results

The results of the mediation analysis are given in Table 2. The results with CC analysis had a sample size of 310 (all births) and 297 (term births), which excluded 147 (32%) and 126 (30%) participants due to missing values. The subjects with missing values were included in the imputation analysis using the *mice* package in RStudio (Stef van Buuren & Groothuis-Oudshoorn, 2011a). Thus the sample sizes increased to 457 and 423 for all births and term births, respectively. Consequently, when comparing MICE(ML) versus CC in Table 2, we saw an overall pattern where the estimates of the regression coefficients θ_1, θ_2, and β_1 became more precise with narrower 95% CIs.

In Table 2, the symbol θ_1 indicates the DE, which is the effect of the air purifier use on the birth weight conditional on controlled blood cadmium level and other covariates. For example, keeping cadmium level and all the covariates constant, the MICE(ML) analysis suggests that intervention was directly associated with a 52 g (95% CI: -24, 129 g) increase in average birth weight in term births. The symbol θ_2 indicates the effect of cadmium level on the birth weight, whereas the symbol β_1 represents the effect of intervention on the cadmium. The estimated parameter values are consistent with the previous findings that air purifier use reduced average blood cadmium concentration (Barn, Gombojav, Ochir, Laagan, et al., 2018). The product of β_1 and θ_2 is the IE, which describes the effect of the HEPA purifier use on the birth weight that passes through the blood cadmium concentration. After applying MICE(ML), the estimated IE were large in magnitude and precise in both all births (17 g, 95% CI: 3, 39 g) and term births (18 g, 95% CI: 3, 40 g), indicating the potential mediating role of the blood cadmium concentration. The PM was calculated by dividing the IE by the TE, and it can be interpreted as the proportion of the effect of HEPA purifiers use on birth weight mediated by the blood cadmium concentration. The PM was 31% (95% CI: -368%, 421%) in all births and 25% (95% CI: -65%, 174%) in term births.

3.4.3 Intention-to-treat analysis of all births

Thus far, the mediation analysis has examined the role of missing data in the mediator variable. However, as indicated in Figure 1, an additional source of missing data in the UGAAR data is the pregnancy losses. There were 47 pregnancy losses due to spontaneous abortion or stillbirth, and these pregnancies were excluded from the mediation analysis because we did not have data on birth weight. However, restricting the analysis to live births can cause selection bias (Chiu et al., 2020). To preserve the benefits of randomization, we conducted an intention-to-treat (ITT) analysis that incorporated the pregnancy losses into the analysis (i.e., incorporated pregnancies with missing values in the outcome).

Given the fact that pregnancy loss shares common causes with preterm birth, such as birth defects and maternal stress (Stein, Susser, Warburton, Wittes, & Kline, 1975), we conducted an ITT analysis assuming that pregnancy losses would have resulted in preterm births if instead they had been born alive. Therefore, preterm status was used as an auxiliary variable in the imputation model in the ITT analysis. Using the gestational age variable, we distinguished between later preterm (35 to 37 weeks) versus moderate preterm (less than 35 weeks). We pessimistically assumed all 47 pregnancy losses would be moderate preterm births.

The results of the ITT analysis are given in the Table 2. The ITT analysis enlarged the sample size to the entire cohort (n = 504) by imputing the birth weight of the 47 pregnancy losses, assuming that all pregnancy losses would have resulting a preterm birth. Then we conducted the same mediation analysis on this full cohort using CC analysis and MICE(ML). Since all the subjects with missing values would be omitted in the CC analysis, the dataset left in the analysis are exactly those who have completed data on all variables (n = 310). Thus, the CC analysis of this full cohort ITT analysis is simply a repeat of the CC analysis of all births and therefore, it was omitted from the table. With MICE(ML), missing birth weights and all the other missing values were imputed.

The results of the ITT analysis in Table 2 shows a similar pattern as the MICE(ML) analysis in all births with slightly wider 95% CIs for target parameters. Including those presumed moderate preterm births, we found a significant indirect effect of 19 g (95% CI: 4, 42 g) and it explained 32% (95% CI: -349%, 410%) of the total effect of HEPA purifier use on birth weight. Thus, incorporating pregnancy losses into the analysis had only a small impact on the results. This may be because the pregnancy losses only occupy less than 10% of the total sample size.

3.5 Simulation study

We conducted a simulation study to evaluate the performance of the proposed MICE(ML) in mediation analysis with missing data, compared with CC analysis, and additionally, compared with a gold-standard analysis that uses the full dataset without any missing data.

3.5.1 Simulation design

In this simulation, the exposure x, mediator m, outcome y, and covariates c_1, c_2, c_3 were randomly generated to mimic a similar but simpler situation than the UGAAR study. The three covariates were considered as mothers' BMI (continuous), child sex (binomial with $p = 0.5$), and mothers' smoking status (binomial with $p = 0.1$), respectively. The quantity x was a binomial random variable with $p = 0.5$, representing the intervention status. The quantity m was a continuous mediator (e.g., log-transformed cadmium) that was assumed to depend on the exposure and mothers' smoking status. The quantity y was birth weight. The two assumed true underlying models for the outcome and mediator were given as:

$$\hat{y} = 3500 + 100\hat{x} - 50\hat{m} + 10\hat{c}_1 - 100\hat{c}_2 - 200\hat{c}_3 + \epsilon$$

$$\hat{m} = -2.5 - 0.2\hat{x} + 0.1\hat{c}_3 + \epsilon.$$

Therefore, the true value of the population parameters $\theta_1, \theta_2,$ and β_1 were 100, -50, and -0.2, respectively. In this setting, the true direct effect of x was 100, indirect effect for m was $(-50) \times (-0.2) = 10$, total effect is $100 + 10 = 110$, and the PM was $10 \div 110 = 0.09$. To mimic the missing data in the UGAAR study, we generated $n = 500$ observations independently and randomly created some missing cells. To make the missingness closer to the reality, we included four different missing patterns: 1) only the mediator has missing values, 2) only c_1 (BMI) has missing values, 3) only c_3 (smoking) has missing values, and 4) both the mediator and c_3 have missing values. In total, 30% of the study participants had missingness in at least one variable. The probability of missingness of each variable was related to the observed values of other variables in the dataset, making the missing mechanism MAR. For example, participants were more reluctant to provide blood sample if they were assigned into the control group.

We randomly generated 1,000 datasets by following the procedures above. We analyzed each dataset using 3 different methods: 1) MICE(ML), 2) CC analysis, and 3) a “gold-standard” analysis. The gold-standard analysis was mediation analysis from fitting Equation 1 and 2 to the entire set of simulated data before generating any missing values. We call it the gold-standard analysis because it entails the best possible scenario, namely the total absence of missing data. In contrast, the CC analysis only analyzed the observed data after generating missing values, whereas the MICE(ML) analyzed the observed data and the imputed data. Thus, the gold-standard analysis represented a benchmark with which to measure the performance improvement of MICE(ML) versus CC analysis. The numbers of bootstrap resampling and MLE simulations were set to 5,000. For each method, we report the average bias estimates, coverage probabilities, and 95% CI ranges from the analysis of the 1,000 simulated datasets.

3.5.2 Simulation results

Table 3 shows the performance of MICE(ML), CC analysis, and gold-standard, and additionally, the true parameters that were used to generate the simulated data. The first section shows the average amount of bias incurred by the three methods, where smaller bias indicates that the point estimates are less biased. Considering the scale of the true parameter, all three methods perform well in point estimates. The CC analysis showed surprisingly good performance, and we suspect that the bootstrap resampling plays an important role in improving its accuracy. We wish we would be able to conduct a simulation study showing more biased estimates in the CC analysis. However, due to the difficulty to run through the simulation codes of the CC analysis, we were not able to achieve this goal. Nonetheless, MICE(ML) showed better performance than the CC analysis in other parts.

In Table 3, the coverage probability indicates the proportion of times the 95% confidence interval contained the true parameters. All methods showed similarly good coverages close to the nominal value of 0.95. Notably, MICE(ML) showed its advantage in the 95% CI width. The 95% CIs were smaller (i.e., more precise) in MICE(ML) analysis than CC analysis. As expected, the CC analysis produced wider confidence intervals because it utilized smaller sample size that discarded the missing data. In contrast, MICE(ML) analysis showed a better performance such that its 95% CIs width were comparable to 95% CIs generated from the gold-standard analysis. The gold-standard analysis generally showed the best performance because it used the entire dataset with no missing data. Overall, we cannot say MICE(ML) is the best approach, but it yields relatively good estimates in the event of missing data and shorter 95% CIs to improve the precision of the estimates.

3.6 Discussion

Mediation analysis has been largely used in health sciences and psychological research. The goal of such an analysis is to examine whether the intermediate variable lies in the causal pathway between the exposure and outcome. In this paper, we extend the general mediation analysis to the context with missing data. We proposed a new approach MICE(ML) that combines ML with MICE. Conceptually, our MICE(ML) method is similar to the bootstrapping method of Wu & Jia (Wu & Jia, 2013) and the *amelidiate* function.(Tingley et al., 2019) However, our approach is more flexible to handle different missing data types. It is less computationally intensive because it incorporates the large sample normal distribution of the ML estimators for parameters in Equations [1] and [2] in order to model uncertainty and calculate 95% CIs. Rather than bootstrapping the data, MICE(ML) simulates from a normal distribution. The whole procedure, including multiple imputation, generating simulations with MLEs and mediation analysis can be finished within several minutes. In contrast, the traditional methods with bootstrap resampling can take hours. The simulation study further demonstrates the good performance of MICE(ML), which was comparable to the gold-standard analysis. With MICE(ML), the mediation output includes point and interval estimates of every single regression coefficient so that it can be used to extend from the application in this study, for example, if researchers want to calculate other types of effect size measure.

We applied this approach to data from the UGAAR study, and we explored the mediating role of blood cadmium concentration in the relationship between HEPA purifier use during pregnancy and infant birth weight in the UGAAR study. We found that MICE(ML) improved the precision of the associations in the mediation analysis because it incorporated the missing data. The 95% CIs of the mediation parameters were narrowed. For example, in Table 2 in the analysis of live births, comparing MICE(ML) with CC analysis, the 95% CI for the direct effect decreased from 37 g (95% CI: -56, 126 g) to 52g (95% CI: -24, 129 g), which was a 16% decrease. This is as important as enlarging the sample sizes. For example, a 16% narrower interval is equivalent to increasing the sample size of the data by 19%. To see why this is the case, note that the standard error of the regression coefficient in linear regression decreases at a rate of $1/\sqrt{n}$. Therefore, a 16% reduction in the standard error is roughly equivalent to multiplying the sample size of the data by a factor of $1/0.84 = 1.19$. Thus, in the UGAAR study, MICE(ML) appeared to recover nearly two thirds of the information that was lost through missing data. In the simulation study, MICE(ML) also showed an improved performance with shorter 95% CIs than CC analysis. The reductions of CI widths ranged from 9% to as large as 42%. The MICE(ML) analysis helped to explain the potential important role of blood cadmium concentration. It suggests that the relationship between HEPA purifier use and birth weight is in part mediated by a change in blood cadmium concentration. The mediated effects explained 31% and 25% of the total effect of the HEPA purifier use during pregnancy on infant birth weight in all births and term births, respectively.

There are some limitations of this study. First, the PM has been criticized as a measure of effect. Although the PM can capture the importance of the mediation pathway, it has been shown to be a highly variable measure (Preacher & Kelley, 2011). The confidence intervals of PM obtained by bootstrapping, maximum likelihood, or other methods can be very wide that exceed ±1.0. Some authors recommended to simply use the confidence interval for the indirect effect (Preacher & Kelley, 2011). The PM has also been criticized based on its interpretation. The quantity $\beta_1\theta_2/(\theta_1 + \beta_1\theta_2)$ can exceed 1.0 (e.g., when θ_1 and $\beta_1\theta_2$ are in opposite directions) or it be negative, depending on the relation of θ_1 to $\beta_1\theta_2$. Given that it is not appropriately interpreted as a proportion, it is less useful than its label implies (Preacher & Kelley, 2011). However, it is the most widely used effect size measure in the literature (Preacher & Kelley, 2011) and a good measurement in the UGAAR study because the point estimates of total effect and the indirect effect are in the same direction. Other available quantitative effect size measures include ratio mediated (Freedman, 2001; Mackinnon, Warsi, & Dwyer, 1995), partially or completely standardized IE (Cheung, 2009), and R^2 measurement (Fairchild, Mackinnon, Taborga, & Taylor, 2009).

Second, although we adjusted for confounders in our regression models, there might exist further unmeasured mediator-outcome confounders. For example, diet during pregnancy influences both blood cadmium concentration (Järup & Åkesson, 2009) and birth weight (Stephenson & Symonds, 2002). Studies have shown that food consumption is a main source of blood cadmium and foods with the highest cadmium content include leafy vegetables, grains, shellfish, and organ meats (Järup & Åkesson, 2009; Wing, Wing, Tidehag, Hallmans, & Sjöström, 1992). However, evidence suggests that the Mongolian populations have low consumption of vegetables and fruits (Public Health Institute, n.d.). Research also shows that the value of cadmium content decreased with increasing animal food consumption (Krajčovičová-Kudláčková et al., 2006). In principle, a sensitivity analysis could be conducted to assess the robustness of our direct and indirect effects to assumptions about unmeasured confounders.

Third, there might be other unmeasured mediators in the pathway from intervention to birth weight. Cadmium may be a marker of sources of environmental toxicants that decrease birth weight. For example, we cannot rule out other coal-fired emissions, such as polycyclic aromatic hydrocarbons (PAHs), as possible mediators of the observed association between intervention and birth weight. Studies have shown that maternal exposure to PAHs have an inverse impact on birth outcomes, especially on birth weight (Anand, Taneja, & Others, 2019; W. A. Jedrychowski et al., 2017; W. Jedrychowski et al., 2003). Coal smoke as a source of cadmium is also a source of PAHs so that the cadmium concentration increase with PAHs concentration (Amster & Levy, 2019; Barn et al., 2019). Particle-bounded PAHs also would likely be affected by HEPA purifier. As a mediator-outcome confounder that is also affected by the exposure (i.e., intervention status), PAHs could be another unmeasured mediator in the pathway from intervention to birth weight.

Although MICE(ML) has its advantage over bootstrap resampling that MICE(ML) is more computationally efficient, the application of MLE has some parametric assumptions (e.g., the point estimates like θ_2 have a normal distribution) (Lehmann, 2009), which may limit the spread of MICE(ML) in other settings, for example, when variable selection is needed. In contrast, the non-parametric bootstrap makes fewer assumptions and allows researchers to estimate the sampling distribution of the point estimates (David, 1966; Obremski & Conover, 1981). However, in some epidemiology studies where variable selection is not the priority, this limitation would not hinder the application of MICE(ML). For example, some demographic variables (e.g., age, gender, income) are needed in the model to adjust for confounding regardless of whether they are significant or not. MLE would be more suitable than bootstrap in these settings.

MICE, as one of the most popular approaches to address missing values has limitations. Its performance has been criticized that MICE perform poorly when there are interactions in the imputation model. Instead, other imputation methods are available to cover this limitation. For example, Conditional Gaussian Model (CGM) has been suggested to reduce bias and produce robust imputations when interactions exist. CGM can impute the categorical variables with log-linear model and impute the continuous outcomes with a conditional JN outcome (Chen, Xie, & Qian, 2011).

We also would have liked to apply MI(BOOT) and BOOT(MI) to the data. However, due to the uncertain R code of these two methods, we did not do it. In addition, BOOT(MI) is totally impractical because it takes way too long.

Regarding the interpretation of results, even though some of the 95% CIs in our results include zero, we still treat them as important. Recently, more and more scientists have risen up against statistical significance and disputed the use of p-values for interpreting results. A recent article by Amrhein, Greenland and McShane (Amrhein, Greenland, & McShane, 2019) advised that researchers should not dismiss results simply because the p-values is greater than 0.05. Such designations might cause genuine effects to be dismissed. For example, in Table 2, with imputation analysis, the results show that the total effect of HEPA purifier use on birth weight was an increase of 71 g, with a 95% CI of (-7, 150 g). Traditionally, this effect would be deemed as not statistically significant because the CI crosses zero. However, considering the regular birth weight of a new-born baby, 7-gram across the zero line is negligible and this effect estimate should be considered important.

Although it is clear that these findings may not be generalizable to other settings, it does not impede its overall aim of identifying the role of blood cadmium in the link in the mediation chain. Findings in this study will still have important implications for public health and air pollution health impact assessment. It should be noted that this approach is not limited to mediation analysis but can also be applied in any studies with missing values to examine the true associations. As future research, we will extend this approach to mediation models with multiple mediators.

Figure 2: A trial profile corresponding to the randomized UGAAR study

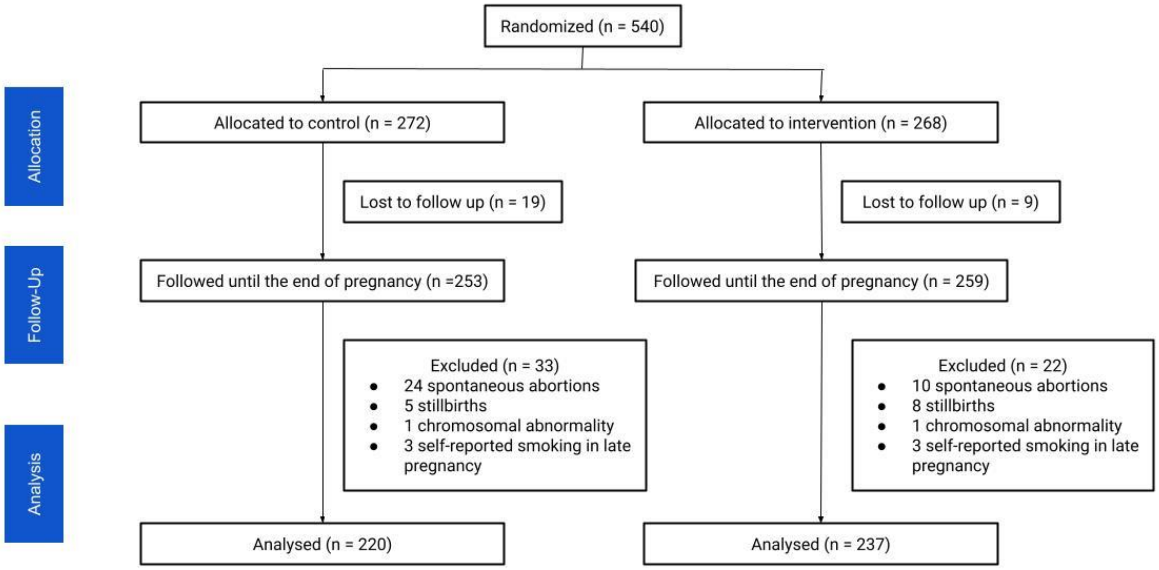

Randomized (n = 540)
Allocation
Allocated to control (n = 272)
Allocated to intervention (n = 268)
Lost to follow up (n = 19)
Lost to follow up (n = 9)
Follow-Up
Followed until the end of pregnancy (n =253)
Followed until the end of pregnancy (n = 259)
Excluded (n = 33)
24 spontaneous abortions
5 stillbirths
1 chromosomal abnormality
3 self-reported smoking in late pregnancy
Excluded (n = 22)
10 spontaneous abortions
8 stillbirths
1 chromosomal abnormality
3 self-reported smoking in late pregnancy
Analysis
Analysed (n = 220)
Analysed (n = 237)

Figure 3: Graphic illustration of the MICE(ML) procedure

m indicates the number of imputed datasets and n indicates the number of simulations. imp = imputed dataset. ψ is a point estimate from the analyses.

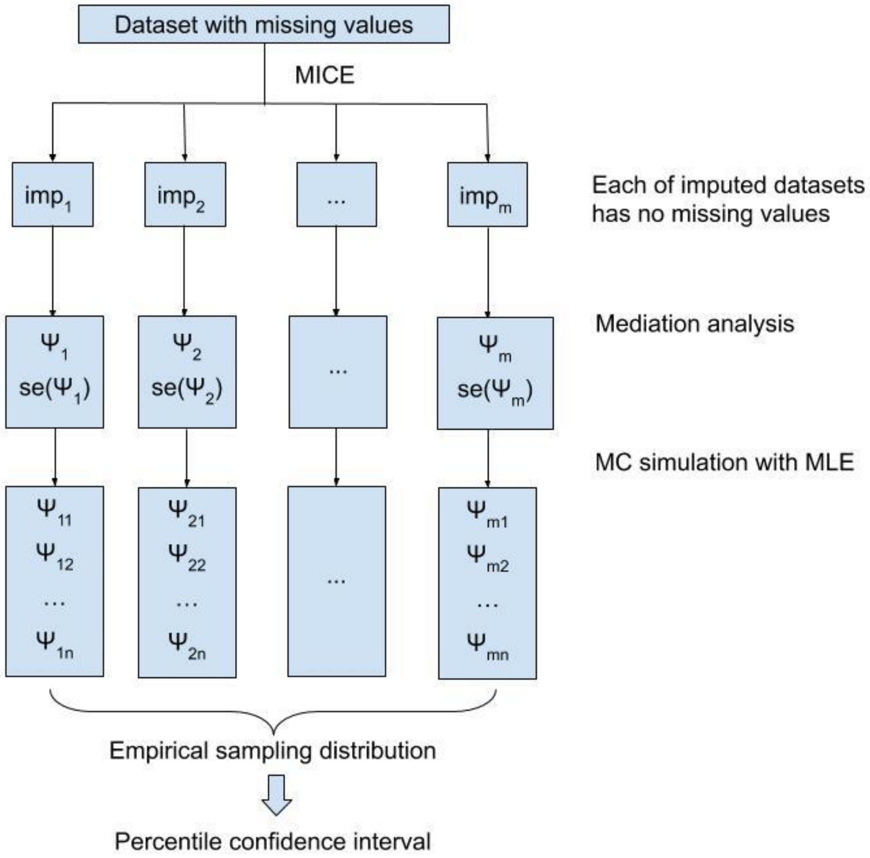

Dataset with missing values
MICE
imp_1
imp_2
...
imp_m
Each of imputed datasets has no missing values
Mediation analysis
Ψ_1 $\text{se}(\Psi_1)$
Ψ_2 $\text{se}(\Psi_2)$
...
Ψ_m $\text{se}(\Psi_m)$
MC simulation with MLE
Ψ_{11} Ψ_{12} … Ψ_{1n}
Ψ_{21} Ψ_{22} … Ψ_{2n}
...
Ψ_{m1} Ψ_{m2} … Ψ_{mn}
Empirical sampling distribution
Percentile confidence interval

Figure 4: A DAG justifying variable selection in the adjusted models among HEPA purifier use, blood cadmium and birth weight

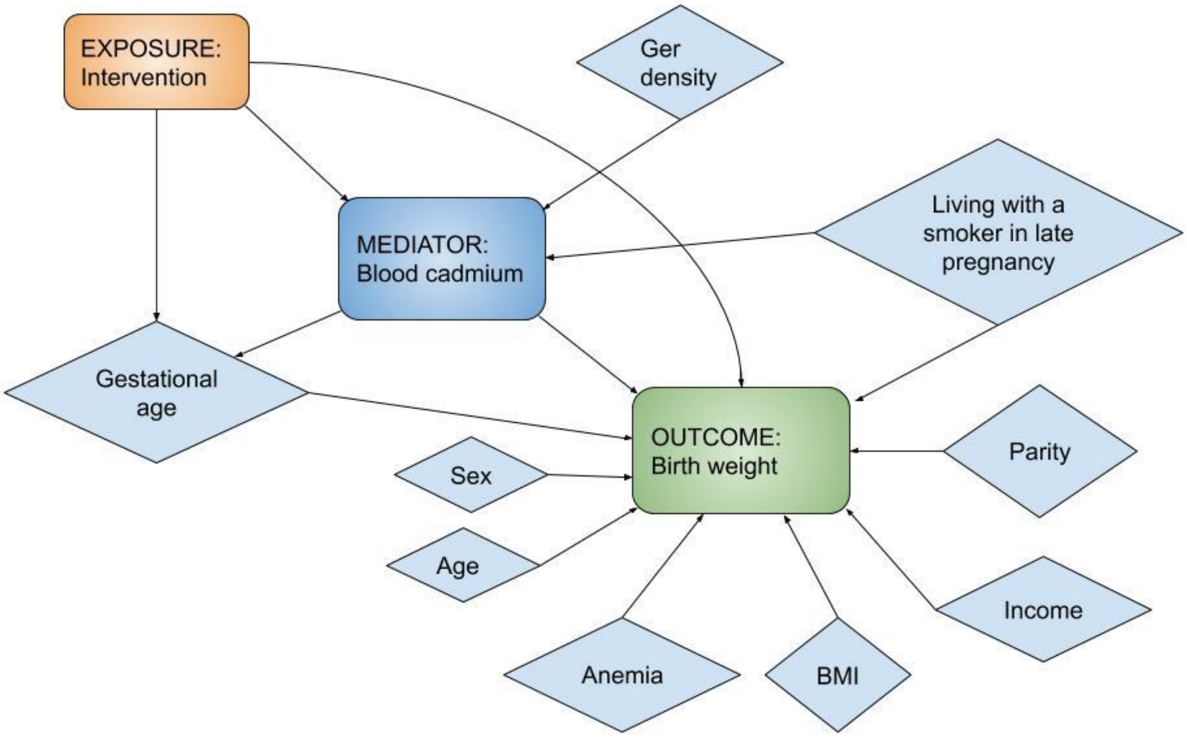


Table 1: Summary of baseline characteristics of control and intervention participants among live births (n = 457) in the UGAAR Study in 2014-15.

Variables	Control (n = 220) Median (25th, 75th percentile) or N (%)	Intervention (n = 237) Median (25th, 75th percentile) or N (%)
Birth outcomes		
Birth weight, g	3450 (3150, 3800)	3550 (3200, 3800)
Preterm birth	10 (5%)	24 (11%)
Mediator		
Blood cadmium concentration, μg/L	0.22 (0.16, 0.31)	0.19 (0.14, 0.27)
Missing	47 (21%)	36 (15%)
Covariates		
Gestational age, weeks	39.5 (38.5, 40.0)	39.5 (38.5, 40.0)
Mother age, years	28 (25, 33)	30 (26, 33)
Pre-pregnancy BMI, kg/m^2	21.6 (19.6, 23.9)	21.4 (19.8, 23.9)
Missing	21 (10%)	8 (3%)
Monthly household income, Tugriks		
< 800,000	77 (35.0%)	90 (38.0%)
≥ 800,000	134 (60.9%)	138 (58.2%)
Missing	9 (4.1%)	9 (3.8%)
Lived with a smoker in late pregnancy	92 (42%)	95 (41%)
Missing	24 (11%)	19 (8%)
Anemia	34 (15%)	53 (43%)
Parity		
0	72 (33%)	79 (33%)

Variables	Control (n = 220) Median (25th, 75th percentile) or N (%)	Intervention (n = 237) Median (25th, 75th percentile) or N (%)
1	87 (40%)	86 (36%)
≥ 2	47 (21%)	61 (26%)
Missing	14 (6%)	11 (5%)
Coal stove density (within 5000m buffer of apartment)	4.04 (3.35, 4.85)	4.23 (3.38, 4.92)
Missing	1 (0%)	3 (1%)
Sex of baby (Male)	122 (51%)	131 (55%)

Table 2: Mediation analysis results comparing CC analysis and MICE(ML) approaches in the UGAAR Study in 2014-15.

	Direct effect of intervention on outcome, θ_1 (g) (95% CI)	Mediator's effect on outcome, θ_2 (g) (95% CI)	Intervention's effect on mediator, β_1 (µg/L) (95% CI)	Indirect effect, $\beta_1\theta_2$ (g) (95% CI)	Total effect, $\theta_1 + \beta_1\theta_2$ (g) (95% CI)	Prop. Mediated, $\beta_1\theta_2/(\theta_1 + \beta_1\theta_2)$ (95% CI)
Live births analysis						
Term births						
CC analysis	37	-91	-0.19	18	55	0.33
(N = 297)	(-56, 126)	(-152, -28)	(-0.38, -0.02)	(1, 42)	(-37, 142)	(-2.48, 3.18)
MICE(ML)	52	-85	-0.23	18	71	0.25
(N = 423)	(-24, 129)	(-137, -31)	(-0.39, -0.06)	(3, 40)	(-7, 150)	(-0.65, 1.74)
All births						
CC analysis	27	-95	-0.17	16	43	0.37
(N = 310)	(-65, 119)	(-157, -31)	(-0.35, 0.00)	(0, 40)	(-47, 134)	(-2.85, 3.75)
MICE(ML)	22	-87	-0.21	17	40	0.31
(N = 457)	(-55, 99)	(-139, -31)	(-0.37, -0.05)	(3, 39)	(-38, 119)	(-3.68, 4.21)
Full cohort ITT analysis						
MICE(ML)	25	-86	-0.24	19	46	0.32
(N = 504)	(-58, 110)	(-142, -31)	(-0.40, -0.08)	(4, 42)	(-40, 131)	(-3.49, 4.10)

Table 3: Bias, coverage probability, and 95% CI width under the simulation study to evaluate the performance of MICE(ML) compared with the gold-standard analysis and the CC analysis.

	True values	Gold-standard	CC	MICE(ML)
	Average bias estimates			
DE, θ_1	100	-0.42	-0.06	-0.08
θ_2	-50	0.45	-0.43	-1.00
β_1	-0.20	0	0	0
IE	10	-0.20	0.01	-0.40
TE	110	-0.62	-0.05	0.11
PM	0.09	0.01	0.01	0.00
	Coverage probability			
DE, θ_1	100	0.952	0.957	0.948
θ_2	-50	0.958	0.952	0.940
β_1	-0.20	0.935	0.938	0.948
IE	10	0.935	0.940	0.947
TE	110	0.949	0.956	0.944
PM	0.09	0.958	0.955	0.946
	95% CI width			
DE, θ_1	100	106	133	107
θ_2	-50	66	80	73
β_1	-0.20	0.28	0.35	0.31
IE	10	20	26	23
TE	110	106	131	109
PM	0.09	0.32	0.57	0.33

Chapter 4 Conclusion

In this thesis, we sought to understand the underlying association between the HEPA purifier use during pregnancy and infant birth weight through blood cadmium concentration. In the presence of missing values, we developed a new approach MICE(ML) to estimate the mediation parameters. We implemented the new approach in the UGAAR study, and we compared the results with the CC analysis, which is the most commonly used approach in the literature (Bouhlila & Sellaouti, 2013). We also conducted a simulation study to evaluate the performance of MICE(ML) side-by-side with gold-standard and CC analysis.

4.1 Contributions

The results of this study are mainly relevant to these stakeholders: environmental health researchers, maternal health researchers, and biostatisticians. The goal of investigating the mediator role of blood cadmium in the association of HEPA purifier use during pregnancy and infant birth weight was also achieved. For example, we found that 31% (95% CI: -368%, 421%) and 25% (95% CI: -65%, 174%) of the effect of HEPA purifier use on birth weight were mediated by the blood cadmium concentration, in all births and term births, respectively. These findings should raise awareness of environmental health and maternal health researchers. The findings may also encourage relevant policies or regulatory regarding chemicals exposures during pregnancy.

Most importantly, the new approach we developed provides a computationally efficient way for biostatisticians to deal with mediation analysis in the presence of missing values. We applied MICE(ML) to the UGAAR study which contains more than 30% of participants with uncomplete data. With the simulation study, we have illustrated the performance of MICE(ML) by comparing its point estimates, coverage probability, and confidence interval width with gold-standard and CC analysis. MICE(ML) accounts for not only the statistical uncertainty from missing data in the imputations but also the uncertainty from sampling fluctuation. In particular, MICE(ML) is much more computationally efficient than existing approaches. This helps the researchers when dealing with big datasets with large percentage of missingness or complex missing patterns. Specifically, MICE(ML) shows best performance in shortening 95% CIs compared to the CC analysis.

MICE(ML) also has its advantage in the flexibility. It can impute missing data in many different data types (e.g., continuous or categorical variables).

4.2 Limitations and future work

While our findings suggest that blood cadmium mediates the effect of HEPA purifier use on birth weight, this study has some limitations. The assumptions of mediation analysis may not be satisfied completely. Although we try to utilize as much information as we have, there might exist some unmeasured mediator-outcome confounders. For example, diet during pregnancy influences both blood cadmium concentration(Järup & Åkesson, 2009) and birth weight (Stephenson & Symonds, 2002). In addition, there might be other unmeasured mediators in the pathway from intervention to birth weight. Cadmium may be a marker of sources of environmental toxicants that decrease birth weight. For example, we cannot rule out other coal-fired emissions, such as polycyclic aromatic hydrocarbons (PAHs), as possible mediators of the observed association between intervention and birth weight.

Regardless, this analysis and simulation study was well designed and conducted. Its findings illustrated the importance of blood cadmium in the causal pathway between HEPA purifier use and infant birth weight. Although it is clear that these findings may not be generalizable to other settings, such as other populations with different exposures, it does not impede its overall aim of identifying the role of blood cadmium in the link in the mediation chain. Findings in this study will still have important implications for public health and air pollution health impact assessment. It suggests the need for future investigations of other inverse health impacts by cadmium exposure during pregnancy. As well, it should be noted that this approach is not limited to mediation analysis but can also be applied in any studies with missing values to examine the true associations. As future research, a promising avenue would be extending this approach to mediation models with multiple mediators. For example, in the UGAAR study we would examine the mediating role of stress on child health outcomes.

Reference

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307.

Amster, E., & Levy, C. L. (2019). Impact of Coal-fired Power Plant Emissions on Children's Health: A Systematic Review of the Epidemiological Literature. *International Journal of Environmental Research and Public Health*, Vol. 16, p. 2008. doi:10.3390/ijerph16112008

Anand, M., Taneja, A., & Others. (2019). Maternal exposure to polycyclic aromatic hydrocarbons (PAHs) exposure and its impact on anthropometric measures of neonates. *Environmental Epidemiology*, *3*, 4.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49.

Barn, P., Gombojav, E., Ochir, C., Boldbaatar, B., Beejin, B., Naidan, G., … Allen, R. W. (2018). The effect of portable HEPA filter air cleaner use during pregnancy on fetal growth: The UGAAR randomized controlled trial. *Environment International*, *121*(Pt 1), 981–989.

Barn, P., Gombojav, E., Ochir, C., Boldbaatar, B., Beejin, B., Naidan, G., … Allen, R. W. (2019). Coal smoke, gestational cadmium exposure, and fetal growth. *Environmental Research*, *179*(Pt B), 108830.

Barn, P., Gombojav, E., Ochir, C., Laagan, B., Beejin, B., Naidan, G., … Allen, R. W. (2018). The effect of portable HEPA filter air cleaners on indoor PM2.5 concentrations and second hand tobacco smoke exposure among pregnant women in Ulaanbaatar, Mongolia: The

UGAAR randomized controlled trial. *The Science of the Total Environment*, *615*, 1379–1389.

Baron, R. M., & Kenny, D. A. (1986). The moderator--mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Bernstein, I. M., Mongeon, J. A., Badger, G. J., Solomon, L., Heil, S. H., & Higgins, S. T. (2005). Maternal smoking and its association with birth weight. *Obstetrics and Gynecology*, *106*(5 Pt 1), 986–991.

Bind, M.-A., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A., Coull, B. A., … Schwartz, J. (2014). Air pollution and gene-specific methylation in the Normative Aging Study: association, effect modification, and mediation analysis. *Epigenetics: Official Journal of the DNA Methylation Society*, *9*(3), 448–458.

Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651–675.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115.

Bonnert, M., Olén, O., Bjureberg, J., Lalouni, M., Hedman-Lagerlöf, E., Serlachius, E., & Ljótsson, B. (2018). The role of avoidance behavior in the treatment of adolescents with irritable bowel syndrome: A mediation analysis. *Behaviour Research and Therapy*, *105*, 27–35.

Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in TIMSS: a case study. *Large-Scale Assessments in Education*, Vol. 1. doi:10.1186/2196-0739-1-4

Brand, J. J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*.

Brand, J., van Buuren, S., le Cessie, S., & van den Hout, W. (2019). Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine*, *38*(2), 210–220.

Carpenter, J. R. (2013). *Multiple imputation and its application*. Chichester, West Sussex: John Wiley & Sons.

Chen, H. Y., Xie, H., & Qian, Y. (2011). Multiple imputation for missing values through conditional Semiparametric odds ratio models. *Biometrics*, *67*(3), 799–809.

Cheung, M. W.-L. (2009). Comparison of methods for constructing confidence intervals of standardized indirect effects. *Behavior Research Methods*, *41*(2), 425–438.

Chiu, Y.-H., Stensrud, M. J., Dahabreh, I. J., Rinaudo, P., Diamond, M. P., Hsu, J., … Hernán, M. A. (2020). The Effect of Prenatal Treatments on Offspring Events in the Presence of Competing Events: An Application to a Randomized Trial of Fertility Therapies. *Epidemiology* , *31*(5), 636–643.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (pp. 412–414). Hillsdale, NJ: Lawrence Erlbaum Associates.

David, H. A. (1966). Handbook of nonparametric statistics. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *8*(3), 553–554.

Desai, M., Kubo, J., Esserman, D., & Terry, M. B. (2011). The handling of missing data in molecular epidemiology studies. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *20*(8), 1571–1579.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.

Fairchild, A. J., Mackinnon, D. P., Taborga, M. P., & Taylor, A. B. (2009). R2 effect-size measures for mediation analysis. *Behavior Research Methods*, *41*(2), 486–498.

Freedman, L. S. (2001). Confidence intervals and statistical power of the ‘Validation’ ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference*, *96*(1), 143–153.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science: The Official Journal of the Society for Prevention Research*, *8*(3), 206–213.

Hafeman, D. M., & Schwartz, S. (2009). Opening the Black Box: a motivation for the assessment of mediation. *International Journal of Epidemiology*, *38*(3), 838–845.

Hoven, H., & Siegrist, J. (2013). Work characteristics, socioeconomic position and health: a systematic review of mediation and moderation effects in prospective studies. *Occupational and Environmental Medicine*, *70*(9), 663–669.

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. *Statistical Strategies for Small Sample Research*, *1*, 195–222.

Hu, J. M. Y., Zhuang, L. H., Bernardo, B. A., & McCandless, L. C. (2018). Statistical challenges in the analysis of biomarkers of environmental chemical exposures for perinatal epidemiology. *Current Epidemiology Reports*. doi:10.1007/s40471-018-0156-x

Järup, L., & Åkesson, A. (2009). Current status of cadmium as an environmental health problem. *Toxicology and Applied Pharmacology*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0041008X09001690

Jedrychowski, W. A., Majewska, R., Spengler, J. D., Camann, D., Roen, E. L., & Perera, F. P. (2017). Prenatal exposure to fine particles and polycyclic aromatic hydrocarbons and birth outcomes: a two-pollutant approach. *International Archives of Occupational and Environmental Health*, *90*(3), 255–264.

Jedrychowski, W., Whyatt, R. M., Camann, D. E., Bawle, U. V., Peki, K., Spengler, J. D., … Perera, F. F. (2003). Effect of prenatal PAH exposure on birth outcomes and neurocognitive development in a cohort of newborns in Poland. Study design and preliminary ambient data. *International Journal of Occupational Medicine and Environmental Health*, *16*(1), 21–29.

Johnson, D. R., & Young, R. (2011). Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family Counseling*, *73*(5), 926–945.

Judd, C. M., & Kenny, D. A. (1981). Process Analysis: Estimating Mediation in Treatment Evaluations. *Evaluation Review*, *5*(5), 602–619.

Kennickell, A. B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, *1*, 41.

Klumparendt, A., Nelson, J., Barenbrügge, J., & Ehring, T. (2019). Associations between childhood maltreatment and adult depression: a mediation analysis. *BMC Psychiatry*, *19*(1), 36.

Krajčovičová-Kudláčková, M., Ursínyová, M., Mašánová, V., Béderová, A., Valachovičová, M., & Others. (2006). Cadmium blood concentrations in relation to nutrition. *Central European Journal of Public Health*, *14*(3), 126–129.

Lall, R. (2016). How Multiple Imputation Makes a Difference. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *24*(4), 414–433.

Lee, H., Herbert, R. D., & McAuley, J. H. (2019). JAMA Guide to Statistics and Methods: Mediation Analysis. *JAMA*, *321*(7), 697–698.

Lee, M., Rahbar, M. H., Brown, M., Gensler, L., Weisman, M., Diekman, L., & Reveille, J. D. (2018). A multiple imputation method based on weighted quantile regression models for longitudinal censored biomarker data with missing values at early visits. *BMC Medical Research Methodology*, *18*(1), 8.

Lehmann, E. L. (2009). Parametric versus nonparametrics: two alternative methodologies. *Journal of Nonparametric Statistics*, *21*(4), 397–405.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

MacKinnon, D. (2012). *Introduction to statistical mediation analysis*. Routledge.

Mackinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research*, *39*(1), 99.

Mackinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A Simulation Study of Mediated Effect Measures. *Multivariate Behavioral Research*, *30*(1), 41.

McDonald, R. P. (1997). Haldane’s Lungs: A Case Study in Path Analysis. *Multivariate Behavioral Research*, *32*(1), 1–38.

Mfutso-Bengo, J., Masiye, F., Molyneux, M., Ndebele, P., & Chilungo, A. (2008). Why do people refuse to take part in biomedical research studies? Evidence from a resource-poor area. *Malawi Medical Journal: The Journal of Medical Association of Malawi*, *20*(2), 57–63.

Morgan, A. J., Mackinnon, A. J., & Jorm, A. F. (2013). Behavior change through automated e-mails: mediation analysis of self-help strategy use for depressive symptoms. *Behaviour Research and Therapy*, *51*(2), 57–62.

Murphy, J., Shevlin, M., Houston, J., & Adamson, G. (2012). Sexual abuse, paranoia, and psychosis: A population-based mediation analysis. *Traumatology*, *18*(1), 37–44.

Nasiri, K., Moodie, E. E. M., & Abenhaim, H. A. (2020). To what extent is the association between race and fetal growth restriction explained by adequacy of prenatal care? A mediation analysis of a retrospective cohort. *American Journal of Epidemiology*. doi:10.1093/aje/kwaa054

Obremski, T. E., & Conover, W. J. (1981). Practical Nonparametric Statistics. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *23*(4), 415.

Pearl, J. (2001). Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420. Presented at the Seattle, Washington. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pearl, J. (2012). The causal mediation formula--a guide to the assessment of pathways and mechanisms. *Prevention Science: The Official Journal of the Society for Prevention Research*, *13*(4), 426–436.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *36*(4), 717–731.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93–115.

Public Health Institute. (n.d.). Third national STEPS Survey on the Prevalence of Noncommunicable Disease and Injury Risk Factors-2013. Retrieved from World Health

Organization website: https://www.who.int/ncds/surveillance/steps/Mongolia_2013_STEPS_Report.pdf

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* , *3*(2), 143–155.

Romano, M. E., Enquobahrie, D. A., Simpson, C., Checkoway, H., & Williams, M. A. (2016). Maternal body burden of cadmium and offspring size at birth. *Environmental Research*, *147*, 461–468.

Royston, P. (2004). Multiple Imputation of Missing Values. *The Stata Journal*, *4*(3), 227–241.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1996). Multiple Imputation After 18 Years. *Journal of the American Statistical Association*, Vol. 91, p. 473. doi:10.2307/2291635

Rubin, D. B. (2009). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation Analysis in Social Psychology: Current Practices and New Recommendations. *Social and Personality Psychology Compass*, *5*(6), 359–371.

SAS Institute. (n.d.). SAS/STAT 14.1 User’s Guide The MI Procedure. Retrieved from https://support.sas.com/documentation/onlinedoc/stat/141/mi.pdf

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, *7*(4), 422–445.

Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, *13*, 290–312.

Springer, M. D., & Thompson, W. E. (1966). The Distribution of Products of Independent Random Variables. *SIAM Journal on Applied Mathematics*, *14*(3), 511–526.

Stata. (n.d.). Multiple Imputation. Retrieved from https://www.stata.com/features/multiple-imputation/

Stein, Z., Susser, M., Warburton, D., Wittes, J., & Kline, J. (1975). Spontaneous abortion as a screening device. The effect of fetal survival on the incidence of birth defects. *American Journal of Epidemiology*, *102*(4), 275–290.

Stephenson, T., & Symonds, M. E. (2002). Maternal nutrition as a determinant of birth weight. *Archives of Disease in Childhood. Fetal and Neonatal Edition*, *86*(1), F4-6.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* , *338*, b2393.

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children’s Mental Health Initiative. *American Journal of Epidemiology*, *169*(9), 1133–1139.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, Vol. 59, pp. 1–38. Retrieved from http://www.jstatsoft.org/v59/i05/

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., Imai, K., Trinh, M., & Wong, W. (2019). *Package “mediation”* (Version 4.5.0). doi:10.1214/10

Travis, C. C. (Ed.). (1993). *Use of Biomarkers in Assessing Health and Environmental Impacts of Chemical Pollutants*. Springer, Boston, MA.

UNICEF. (n.d.). Environment & air pollution. Retrieved from UNICEF Mongolia website: https://www.unicef.org/mongolia/environment-air-pollution

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681–694.

Van Buuren, S., & Oudshoorn, K. (2000). Multivariate Imputation by Chained Equations: MICE V1. 0 User’s Manual, volume Pg/Vgz/00.038. *TNO Prevention and Health, Leiden. URL Http://Www. Stefvanbuuren. Nl/Publications/Mice% 20v1. 0% 20manual% 20tno00038*, *202000*.

van Buuren, Stef. (2018). *Flexible Imputation of Missing Data, Second Edition*. CRC Press.

van Buuren, Stef, & Groothuis-Oudshoorn, K. (2011a). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, Vol. 45, pp. 1–67. Retrieved from https://www.jstatsoft.org/v45/i03/

van Buuren, Stef, & Groothuis-Oudshoorn, K. (2011b). mice: Multivariate Imputation by Chained Equations inR. *Journal of Statistical Software*, Vol. 45. doi:10.18637/jss.v045.i03 van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to bayesian analysis: applications to developmental research. *Child Development*, *85*(3), 842–860.

Vanderweele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Cary, NC: Oxford University Press.

VanderWeele, T. J. (2016). Mediation Analysis: A Practitioner's Guide. *Annual Review of Public Health*, *37*(1), 17–32.

VanderWeele, T. J., & VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

von Hippel, P., & Lynch, J. (2013). Efficiency Gains from Using Auxiliary Variables in Imputation. Retrieved from http://arxiv.org/abs/1311.5249

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.

Wing, A. M., Wing, K., Tidehag, P., Hallmans, G., & Sjöström, R. (1992). Cadmium accumulation from diets with and without wheat bran in rats with different iron status. *Nutrition Research*, Vol. 12, pp. 1205–1215. doi:10.1016/s0271-5317(05)80777-8

Wu, W., & Jia, F. (2013). A New Procedure to Test Mediation With Missing Data Through Nonparametric Bootstrapping and Multiple Imputation. *Multivariate Behavioral Research*, *48*(5), 663–691.

Zhang, J., Wedel, M., & Pieters, R. (2009). Sales effects of attention to feature advertisements: A Bayesian mediation analysis. *JMR, Journal of Marketing Research*, *46*(5), 669–681.

Zhang, Z., Wang, L., & Tong, X. (2015). Mediation Analysis with Missing Data Through Multiple Imputation and Bootstrap. *Quantitative Psychology Research*, 341–355. Springer International Publishing.

Appendix A. Supplementary Table 1 for Manuscript Chapter 3

To test the performance of MICE(ML), we also conducted another simulation study with simpler missing pattern: only the mediator (m) has missing values. Other factors remained the same, for example, the values of the true parameters, percentage of missingness (i.e., 30%), and 1,000 randomly generated datasets with 500 observations in each dataset. Likewise, we analyzed each dataset using 3 different methods: 1) MICE(ML), 2) CC analysis, and 3) gold-standard analysis. Supplementary Table 1 shows the performance of these methods. Similarly, the three approaches showed comparable performance in average bias estimates and coverage probability. MICE(ML) showed a little bigger but negligible bias than CC analysis in most of the point estimates.

The difference of the three approaches is evident in the 95% CI width. As shown in the simulation study in Section 3.5, MICE(ML) has a distinct advantage that it produces a much shorter 95% CI width and this width can be comparable to the 95% CI width generated from the gold-standard analysis. This advantage was inherited in this simulation study, as well. The 95% CI width of θ_1, TE and PM produced by MICE(ML) are apparently shorter than the widths produced by the CC analysis (e.g., 108 in MICE(ML) vs. 133 in CC analysis for θ_1) and are very close to the widths produced by the gold-standard analysis (e.g., 108 in MICE(ML) vs. 107 in gold-standard analysis for θ_1). We also note that in this simulation study with a simpler missing pattern, this advantage of shorter 95% CI width looks not as prominent as the result in Table 3, which is the result of the simulation study with more complex missing pattern. It demonstrates that MICE(ML) plays an important role in imputing plausible values and address the imputation uncertainty. In the settings with more complex missing patterns or a larger missing percentage, this benefit over the CC analysis can restore the missing part of the data and increase the significance level of the findings.

Supplementary Table 1 Bias, coverage probability, and 95% CI width under the simulation study with the simpler missing pattern to evaluate the performance of MICE(ML) compared with the gold-standard analysis and the CC analysis.

	True values	Gold-standard	CC	MICE(ML)
		Average bias estimates		
DE, θ_1	100	-0.73	-1.15	1.66
θ_2	-50	0.24	-0.96	-0.66
β_1	-0.20	-0.01	0.00	0.00
IE	10	0.30	-0.10	-0.63
TE	110	-0.42	-1.24	1.70
PM	0.09	0.01	0.01	0.00
		Coverage probability		
DE, θ_1	100	0.958	0.953	0.942
θ_2	-50	0.945	0.955	0.948
β_1	-0.20	0.960	0.955	0.946
IE	10	0.941	0.949	0.954
TE	110	0.955	0.952	0.951
PM	0.09	0.940	0.966	0.945
		95% CI width		
DE, θ_1	100	107	133	108
θ_2	-50	66	80	79
β_1	-0.20	0.28	0.35	0.35
IE	10	20	26	26
TE	110	106	132	109
PM	0.09	0.30	0.61	0.32

Appendix B. R code of the MSc thesis

R code of the mediation analysis with MICE(ML) in the UGAAR study for Manuscript Chapter 3

Preparation work

```
library(dplyr)
library(mice)
library(boot)

indirectsaved = function(
  formula.y,formula.m,
  dataset, random){
  d = dataset[random,]
  model.y = lm(formula.y, data = d)
  model.m = lm(formula.m, data = d)

  theta1 = coef(model.y)[2]
  theta2 = coef(model.y)[3]
  beta = coef(model.m)[2]
  IE = theta2*beta
  TE = IE+theta1
  PM = IE/TE
  return(c(theta1, theta2, beta, IE, TE,PM))
}

n.imputations = 20
n.simulations = 10000
set.seed(13927489)
```

Import UGAAR data

```
allbirth <-read.csv("~/Documents/MSc/UGAAR/data/bwt457.csv")
termbirth <- filter(allbirth, allbirth$ga >= 37)
```

1. **All births**

CC analysis

```
result <- boot(data = allbirth,
               statistic = indirectsaved,
               formula.y = bw ~ Intervention + logcd + BMI_prepreg + age4 + i
ncome4 + anemia + parity + ger_den + lws_late + sex + ga + I(ga^2),
               formula.m = logcd ~ Intervention + BMI_prepreg + age4 + income
4 + anemia + parity + ger_den + lws_late + sex + ga + I(ga^2),
               R = n.simulations)
result
for (i in 1:6) {
  a <- boot.ci(boot.out=result, conf = 0.95, type = "perc", index = i)
  print(a)
}
```

MICE(ML)

```
mi.all.table <- setNames(data.frame(matrix(data = NA, nrow = 6, ncol = 3)),
                    c("point","L","R"))
rownames(mi.all.table) <- c("theta1", "theta2", "beta", "IE", "TE", "PM")

set.seed(13927489)
  # theta1
    simulations.theta1 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

    for (i in 1:n.imputations){
      theta1.hat <- coef(summary(y.MICE.all$analyses[[i]]))[2,1]
      theta1.hat.se <- coef(summary(y.MICE.all$analyses[[i]]))[2,2]
      simulations.theta1[,i] <- rnorm(n.simulations, theta1.hat, theta1.hat.s
e)
    }
    mi.all.table[1,1] <- quantile(simulations.theta1, probs=0.5)
    # 95% confidence interval
    mi.all.table[1,c(2,3)] <- quantile(simulations.theta1, probs=c(0.025, 0.9
75))

  # theta2
    simulations.theta2 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

    for (i in 1:n.imputations){
      theta2.hat <- coef(summary(y.MICE.all$analyses[[i]]))[3,1]
      theta2.hat.se <- coef(summary(y.MICE.all$analyses[[i]]))[3,2]
      simulations.theta2[,i] <- rnorm(n.simulations, theta2.hat, theta2.hat.s
e)
    }
    mi.all.table[2,1] <- quantile(simulations.theta2, probs=0.5)
    mi.all.table[2,c(2,3)] <- quantile(simulations.theta2, probs=c(0.025, 0.9
75))
  # beta
    simulations.beta <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)
    for (i in 1:n.imputations){
      beta.hat <- coef(summary(m.MICE.all$analyses[[i]]))[2,1]
      beta.hat.se <- coef(summary(m.MICE.all$analyses[[i]]))[2,2]
      simulations.beta[,i] <- rnorm(n.simulations, beta.hat, beta.hat.se)
    }
    mi.all.table[3,1] <- quantile(simulations.beta, probs=0.5)
    mi.all.table[3,c(2,3)] <- quantile(simulations.beta, probs=c(0.025, 0.975
))

  # IE
  mi.all.table[4,1] <- quantile(simulations.theta2*simulations.beta, probs=0.
5)

  mi.all.table[4,c(2,3)] <- quantile(simulations.theta2*simulations.beta, pro
bs=c(0.025, 0.975))

  # TE
  mi.all.table[5,1] <- quantile(simulations.theta1 + simulations.theta2*simul
ations.beta, probs=0.5)
  mi.all.table[5,c(2,3)] <- quantile(simulations.theta1 + simulations.theta2*
simulations.beta, probs=c(0.025, 0.975))

  # PM
  mi.all.table[6,1] <- quantile( (simulations.theta2*simulations.beta)/(simul
ations.theta1 + simulations.theta2*simulations.beta), probs=0.5)
  mi.all.table[6,c(2,3)] <- quantile( (simulations.theta2*simulations.beta)/(
simulations.theta1 + simulations.theta2*simulations.beta), probs=c(0.025, 0.9
75))

round(mi.all.table,2)
```

2. Term births

CC analysis

```
set.seed(13927489)

result <- boot(data = termbirth,
               statistic = indirectsaved,
               formula.y = bw ~ Intervention + logcd + BMI_prepreg + age4 + i
ncome4 + anemia + parity + ger_den + lws_late + sex + ga + I(ga^2),
               formula.m = logcd ~ Intervention + BMI_prepreg + age4 + income
4 + anemia + parity + ger_den + lws_late + sex + ga + I(ga^2),
               R = n.simulations)
result
for (i in 1:6) {
  a <- boot.ci(boot.out=result, conf = 0.95, type = "perc", index = i)
  print(a)
}
```

MICE(ML)

```
mi.term.table <- setNames(data.frame(matrix(data = NA, nrow = 6, ncol = 3)),
                          c("point","L","R"))
rownames(mi.term.table) <- c("theta1", "theta2", "beta", "IE", "TE", "PM")

MI.term <- mice(termbirth, m = n.imputations, printFlag = F, maxit = 5, seed
= 123)

y.MICE.term <- with(MI.term, lm(bw ~ Intervention + logcd + age4 + income4 +
BMI_prepreg + parity + anemia + lws_late + ger_den + sex + ga + I(ga^2)))
summary(pool(y.MICE.term))[c(2,3),]

m.MICE.term <- with(MI.term, lm(logcd ~ Intervention + age4 + income4 + BMI_p
repreg + parity + anemia + lws_late + ger_den + sex + ga + I(ga^2)))

set.seed(13927489)
  # theta1
    simulations.theta1 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

    for (i in 1:n.imputations){
      theta1.hat <- coef(summary(y.MICE.term$analyses[[i]]))[2,1]
      theta1.hat.se <- coef(summary(y.MICE.term$analyses[[i]]))[2,2]
      simulations.theta1[,i] <- rnorm(n.simulations, theta1.hat, theta1.hat.s
e)
    }
    mi.term.table[1,1] <- quantile(simulations.theta1, probs=0.5)
    # 95% confidence interval
    mi.term.table[1,c(2,3)] <- quantile(simulations.theta1, probs=c(0.025, 0.
975))

  # theta2
    simulations.theta2 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

    for (i in 1:n.imputations){
      theta2.hat <- coef(summary(y.MICE.term$analyses[[i]]))[3,1]
      theta2.hat.se <- coef(summary(y.MICE.term$analyses[[i]]))[3,2]
      simulations.theta2[,i] <- rnorm(n.simulations, theta2.hat, theta2.hat.s
e)
    }
     mi.term.table[2,1] <- quantile(simulations.theta2, probs=0.5)
     mi.term.table[2,c(2,3)] <- quantile(simulations.theta2, probs=c(0.025, 0
.975))
  # beta
    simulations.beta <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)
    for (i in 1:n.imputations){
      beta.hat <- coef(summary(m.MICE.term$analyses[[i]]))[2,1]
      beta.hat.se <- coef(summary(m.MICE.term$analyses[[i]]))[2,2]
      simulations.beta[,i] <- rnorm(n.simulations, beta.hat, beta.hat.se)
    }
     mi.term.table[3,1] <- quantile(simulations.beta, probs=0.5)
     mi.term.table[3,c(2,3)] <- quantile(simulations.beta, probs=c(0.025, 0.9
75))

  # IE
  mi.term.table[4,1] <- quantile(simulations.theta2*simulations.beta, probs=0
.5)
  mi.term.table[4,c(2,3)] <- quantile(simulations.theta2*simulations.beta, pr
obs=c(0.025, 0.975))

  # TE
  mi.term.table[5,1] <- quantile(simulations.theta1 + simulations.theta2*simu
lations.beta, probs=0.5)
  mi.term.table[5,c(2,3)] <- quantile(simulations.theta1 + simulations.theta2
*simulations.beta, probs=c(0.025, 0.975))

  # PM
  mi.term.table[6,1] <- quantile( (simulations.theta2*simulations.beta)/(simu
lations.theta1 + simulations.theta2*simulations.beta), probs=0.5)
  mi.term.table[6,c(2,3)] <- quantile( (simulations.theta2*simulations.beta)/
(simulations.theta1 + simulations.theta2*simulations.beta), probs=c(0.025, 0.
975))

  round(mi.term.table,2)
```

3. ITT analysis

```
nobw <- read.csv("~/Documents/MSc/UGAAR/data/abortions_stillbirths.csv")
# combine all births and birth losses
itt <- rbind(nobw, allbirth)
```

CC analysis of ITT is the same as all births

MICE(ML)

```
mi.itt.table <- setNames(data.frame(matrix(data = NA, nrow = 6, ncol = 3)),
                    c("point","L","R"))
rownames(mi.itt.table) <- c("theta1", "theta2", "beta", "IE", "TE", "PM")

set.seed(13927489)
  # theta1
    simulations.theta1 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

    for (i in 1:n.imputations){
      theta1.hat <- coef(summary(y.MICE.itt$analyses[[i]]))[2,1]
      theta1.hat.se <- coef(summary(y.MICE.itt$analyses[[i]]))[2,2]
      simulations.theta1[,i] <- rnorm(n.simulations, theta1.hat, theta1.hat.s
e)
    }
    mi.itt.table[1,1] <- quantile(simulations.theta1, probs=0.5)
    # 95% confidence interval
    mi.itt.table[1,c(2,3)] <- quantile(simulations.theta1, probs=c(0.025, 0.9
75))

  # theta2
    simulations.theta2 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

    for (i in 1:n.imputations){
      theta2.hat <- coef(summary(y.MICE.itt$analyses[[i]]))[3,1]
      theta2.hat.se <- coef(summary(y.MICE.itt$analyses[[i]]))[3,2]
      simulations.theta2[,i] <- rnorm(n.simulations, theta2.hat, theta2.hat.s
e)
    }
    mi.itt.table[2,1] <- quantile(simulations.theta2, probs=0.5)
    mi.itt.table[2,c(2,3)] <- quantile(simulations.theta2, probs=c(0.025, 0.9
75))
  # beta
    simulations.beta <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)
    for (i in 1:n.imputations){
      beta.hat <- coef(summary(m.MICE.itt$analyses[[i]]))[2,1]
      beta.hat.se <- coef(summary(m.MICE.itt$analyses[[i]]))[2,2]
      simulations.beta[,i] <- rnorm(n.simulations, beta.hat, beta.hat.se)
    }
    mi.itt.table[3,1] <- quantile(simulations.beta, probs=0.5)
    mi.itt.table[3,c(2,3)] <- quantile(simulations.beta, probs=c(0.025, 0.975
))

  # IE
  mi.itt.table[4,1] <- quantile(simulations.theta2*simulations.beta, probs=0.
5)
  mi.itt.table[4,c(2,3)] <- quantile(simulations.theta2*simulations.beta, pro
bs=c(0.025, 0.975))

  # TE
  mi.itt.table[5,1] <- quantile(simulations.theta1 + simulations.theta2*simul
ations.beta, probs=0.5)
  mi.itt.table[5,c(2,3)] <- quantile(simulations.theta1 + simulations.theta2*
simulations.beta, probs=c(0.025, 0.975))

  # PM
  mi.itt.table[6,1] <- quantile( (simulations.theta2*simulations.beta)/(simul
ations.theta1 + simulations.theta2*simulations.beta), probs=0.5)
  mi.itt.table[6,c(2,3)] <- quantile( (simulations.theta2*simulations.beta)/(
simulations.theta1 + simulations.theta2*simulations.beta), probs=c(0.025, 0.9
75))
round(mi.itt.table,2)
```

R code of the simulation study for Manuscript Chapter 3

Packages to be installed

```r
library(mice)
library(dplyr)
library(mediation)
```

Some numbers in the simulation

```r
n = 500 # sample size
nsims = 1000 # number of simulations, replicate the whole process 1000 times

miss.prop = 0.3

# MICE(ML)
n.imputations = 20
n.simulations = 5000

set.seed(1837794)

# Self-build functions
## Average bias
bias_calculator <- function(values, mean, nsims) {
  result <- sum(values - mean)/nsims
  return(result)
}

## coverage probability
cov_calculator <- function(flags, nsims) {
    result <- sum(flags == T)/nsims
    return(result)
}
```

Tables to store the results

```r
ind <- c('theta1','theta2','beta','IE','DE','TE','PM')
ind.bias <- c('theta1','theta2','beta','IE','TE','PM')

bias.table <- setNames(data.frame(matrix(data = NA, nrow = 6, ncol = 3)),
                    c('Gold','CC','MICE(ML)'))
    med.bias.table <- setNames(data.frame(matrix(data = NA, nrow = 4, ncol =
3)),
                    c('Gold','CC','MICE(ML)'))
```

```r
coverage.table <- setNames(data.frame(matrix(data = NA, nrow = 7, ncol = 3)),

                    c('Gold','CC','MICE(ML)'))

power.table <- setNames(data.frame(matrix(data = NA, nrow = 7, ncol = 3)),
                    c('Gold','CC','MICE(ML)'))

length.table <- setNames(data.frame(matrix(data = NA, nrow = 7, ncol = 3)),
                    c('Gold','CC','MICE(ML)'))

rownames(bias.table)<- ind.bias
rownames(coverage.table)<- ind
rownames(length.table)<- ind
```

True population parameters

θ_1 = 100, θ_2 = -50, β = -0.2,

IE = 10, DE (θ_1) = 100, TE = 110, PM = 0.09

1. Gold: (randomly generated) sample statistics - full dataset

```r
# Tables to store results
table.gold <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 12))
,
                    c("theta1", "theta1_L","theta1_R",
                      "theta2", "theta2_L","theta2_R",
                      "beta", "beta_L","beta_R",
                      "IE", "TE","PM"))
med.table.gold <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol =
12)),
                           c("MIE","MIE_L","MIE_R",
                             "MDE","MDE_L","MDE_R",
                             "MTE","MTE_L","MTE_R",
                             "MPM","MPM_L","MPM_R"))

for (i in 1:nsims){ print(i)
  # randomly create samples
  c1 <- rnorm(n, 25, 3)
  c2 <- rbinom(n, 1, 0.5)
  c3 <- rbinom(n, 1, 0.1)
  x <- rbinom(n, 1, 0.5)
  m <- rnorm(n, mean=-2.5 + -0.2 * x + 0.1 * c3, sd=0.8)
  y <- rnorm(n, 3500 + 100 * x + -50*m + 10*c1 + -100*c2 + -200*c3, 300)
  gold.data <- data.frame(y, x, m, c1, c2, c3)
  # fit 2 models and the mediation
  y.gold <- lm(y ~ x + m + c1 + c2 + c3, data=gold.data)
  m.gold <- lm(m ~ x + c1 + c2 + c3, data=gold.data)

  # use mediation package to get the interval
  gold.ci <- mediate(m.gold, y.gold, sims = nsims, boot = TRUE, boot.ci.type
= "perc", treat = "x", mediator = "m", conf.level = 0.95)
  med <- summary(gold.ci)

  # fill the table
  table.gold[i,"theta1"] <- coef(summary(y.gold))[2,1]
  table.gold[i,2] <- confint(y.gold, c("x","m"), 0.95)[1,1] # theta1_L
  table.gold[i,3] <- confint(y.gold, c("x","m"), 0.95)[1,2]  # theta1_R

  table.gold[i,"theta2"] <- coef(summary(y.gold))[3,1]
  table.gold[i,5] <- confint(y.gold, c("x","m"), 0.95)[2,1] # theta2_L
  table.gold[i,6] <- confint(y.gold, c("x","m"), 0.95)[2,2] # theta2_R

  table.gold[i,"beta"] <- coef(summary(m.gold))[2,1]
  table.gold[i,8] <- confint(m.gold, "x", 0.95)[1,1] # beta_L
  table.gold[i,9] <- confint(m.gold, "x", 0.95)[1,2] # beta_R

    ## These effects are calculated from regression coefficients.
    ## therefore, it cannot obtain se or CI because of "product"
  table.gold[i,"IE"] <- table.gold[i,"theta2"]*table.gold[i,"beta"]
  table.gold[i,"TE"] <- table.gold[i,"IE"] + table.gold[i,"theta1"]
  table.gold[i,"PM"] <- table.gold[i,"IE"]/table.gold[i,"TE"]

  # fill the med table
    ## These effects are calculated from mediation package
    ## we can obtain CI of IE, DE, TE, PM and check if point estimates by two
 methods are same
  med.table.gold[i,"MIE"] <- med$d0
  med.table.gold[i,"MIE_L"] <- med$d0.ci[[1]]
  med.table.gold[i,"MIE_R"] <- med$d0.ci[[2]]

  med.table.gold[i,"MDE"] <- med$z0
  med.table.gold[i,"MDE_L"] <- med$z0.ci[[1]]
  med.table.gold[i,"MDE_R"] <- med$z0.ci[[2]]

  med.table.gold[i,"MTE"] <- med$tau.coef
  med.table.gold[i,"MTE_L"] <- med$tau.ci[[1]]
  med.table.gold[i,"MTE_R"] <- med$tau.ci[[2]]

  med.table.gold[i,"MPM"] <-  med$n0
  med.table.gold[i,"MPM_L"] <- med$n0.ci[[1]]
  med.table.gold[i,"MPM_R"] <- med$n0.ci[[2]]
}

# table.gold
```

```r
(mean.table.gold <- round(apply(table.gold[,c(1,4,7,10:12)], 2, mean),3))
#med.table.gold
(mean.med.gold<- round(apply(med.table.gold[,c(1,4,7,10)], 2, mean),3))
```

Bias - gold

```r
gold.bias <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 6)),
                    c("theta1","theta2", "beta","IE","TE","PM"))

gold.bias["theta1"] <- bias_calculator(table.gold[,1], 100, nsims)
gold.bias["theta2"] <- bias_calculator(table.gold[,4], -50, nsims)
gold.bias["beta"] <- bias_calculator(table.gold[,7], -0.2, nsims)
gold.bias["IE"] <- bias_calculator(table.gold[,10], 10, nsims)
gold.bias["TE"] <- bias_calculator(table.gold[,11], 110, nsims)
gold.bias["PM"] <- bias_calculator(table.gold[,12], 0.09, nsims)

# just check if they are same
gold.med.bias <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 4)),
                          c("MIE","MDE","MTE","MPM"))
gold.med.bias["MIE"] <- bias_calculator(med.table.gold[,1], 10, nsims)
gold.med.bias["MDE"] <- bias_calculator(med.table.gold[,4], 100, nsims)
gold.med.bias["MTE"] <- bias_calculator(med.table.gold[,7], 110, nsims)
gold.med.bias["MPM"] <- bias_calculator(med.table.gold[,10], 0.09, nsims)


bias.table[,1] <- t(round(gold.bias,2))
bias.table
med.bias.table[,1] <- t(round(gold.med.bias,2))
med.bias.table
# no difference
```

Coverage - gold

```r
# coverage of parameters (theta1, theta2, beta)
gold.coverage <- table.gold %>%
  mutate(theta1_coverage = ifelse(theta1_L <= 100 & theta1_R >= 100,
                              T,
                              F)) %>%
  mutate(theta2_coverage = ifelse(theta2_L <= -50 & theta2_R >= -50,
                              T,
                              F)) %>%
  mutate(beta_coverage = ifelse(beta_L <= -0.2  & beta_R >= -0.2,
                              T,
                              F))
# coverage of effects (IE, DE, TE, PM)
med.gold.coverage <- med.table.gold %>%
  mutate(MIE_coverage = ifelse(MIE_L <= 10 & MIE_R >= 10,
                              T,
                              F)) %>%
  mutate(MDE_coverage = ifelse(MDE_L <= 100 & MDE_R >= 100,
                              T,
                              F)) %>%
  mutate(MTE_coverage = ifelse(MTE_L <= 110  & MTE_R >= 110,
                              T,
                              F)) %>%
  mutate(MPM_coverage = ifelse(MPM_L <= 0.09  & MPM_R >= 0.09,
                              T,
                              F))
gold.coverage.table <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol =
 7)), c("theta1", "theta2","beta","MIE","MDE","MTE","MPM"))

  gold.coverage.table[,1] <- cov_calculator(flags = gold.coverage$theta1_cove
rage, nsims = nsims)
  gold.coverage.table[,2] <- cov_calculator(flags = gold.coverage$theta2_cove
rage, nsims = nsims)
  gold.coverage.table[,3] <- cov_calculator(flags = gold.coverage$beta_covera
ge, nsims = nsims)

  gold.coverage.table[,4] <- cov_calculator(flags = med.gold.coverage$MIE_cov
erage, nsims = nsims)
  gold.coverage.table[,5] <- cov_calculator(flags = med.gold.coverage$MDE_cov
erage, nsims = nsims)
  gold.coverage.table[,6] <- cov_calculator(flags = med.gold.coverage$MTE_cov
erage, nsims = nsims)
  gold.coverage.table[,7] <- cov_calculator(flags = med.gold.coverage$MPM_cov
erage, nsims = nsims)


gold.coverage.table
coverage.table[,1] <- t(gold.coverage.table)
coverage.table
```

Interval Length

```
length.gold <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 7))
,
                    c("theta1","theta2","beta2","IE","DE","TE","PM"))
length.gold[,1] <- abs(table.gold[,3] - table.gold[,2])
length.gold[,2] <- abs(table.gold[,6] - table.gold[,5])
length.gold[,3] <- abs(table.gold[,9] - table.gold[,8])
length.gold[,4] <- abs(med.table.gold[,3] - med.table.gold[,2])
length.gold[,5] <- abs(med.table.gold[,6] - med.table.gold[,5])
length.gold[,6] <- abs(med.table.gold[,9] - med.table.gold[,8])
length.gold[,7] <- abs(med.table.gold[,12] - med.table.gold[,11])

length.gold.table <- round(colMeans(length.gold),2)
length.table[,1] <-length.gold.table
length.table
```

2. CC analysis

```
table.CC <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 12)),
                     c("theta1","theta1_L","theta1_R",
                       "theta2","theta2_L","theta2_R",
                       "beta","beta_L","beta_R",
                       "IE", "TE", "PM"))
med.table.CC <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 12
)),
                         c("MIE","MIE_L","MIE_R",
                           "MDE","MDE_L","MDE_R",
                           "MTE","MTE_L","MTE_R",
                           "MPM","MPM_L","MPM_R"))

for (i in 1:nsims) { print(i)
  c1 <- rnorm(n, 25, 3)
  c2 <- rbinom(n, 1, 0.5)
  c3 <- rbinom(n, 1, 0.1)
  x <- rbinom(n, 1, 0.5)
  m <- rnorm(n, mean=-2.5 + -0.2 * x + 0.1 * c3, sd=0.8)
  y <- rnorm(n, 3500 + 100 * x + -50*m + 10*c1 + -100*c2 + -200*c3, 300)
  gold.data <- data.frame(y, x, m, c1, c2, c3)

  # make some missingness
  mypattern <-  matrix(c(1,1,0,1,1,1, # only m missing
                         1,1,1,0,1,1, # only c1 missing (BMI)
                         1,1,1,1,1,0, # only c3 missing (smoke)
                         1,1,0,1,1,0), 4,6, byrow = T) # m and c3 missing
  myfreq <- c(0.48, 0.2, 0.2, 0.12) # frequency for each pattern
  myweights <- matrix(c(0, 1, 0, 0.1, 0, 0.3,
                        # only m is missing: missingness of m is correlated t
o x, c1, c3
                        # x is weighted 10 times as heavy as c1, etc.

                        0, 1, 0, 0, 0, 0,
                        # only c1 is missing : missingness of BMI is cor to x


                        0, 1, 0, 0.5, 0, 0,
                        # only c3 is missing: missingness of smoke is cor to
x, BMI

                        0, 1, 0, 0, 0.4, 0), 4,6,byrow = T)
                        # m and c3 are missing: missingness of both m & smoke
 are cor to x, BMI
```

```
  miss <- ampute(gold.data,freq = myfreq , patterns = mypattern, weights = my
weights, prop = miss.prop, mech = "MAR")
  observed.data <- miss$amp

  # fit 2 models and the mediation
  y.cc <- lm(y ~ x + m + c1 + c2 + c3, data=observed.data)
  m.cc <- lm(m ~ x + c1 + c2 + c3, data=observed.data)

  # use mediation package to get the interval
  CC.ci <- mediate(m.cc, y.cc, sims = nsims, boot = TRUE, boot.ci.type = "per
c", treat = "x", mediator = "m", conf.level = 0.95)
  med <- summary(CC.ci)

  # fill the table
  table.CC[i,1] <- coef(summary(y.cc))[2,1] # DE = theta1
  table.CC[i,2] <- confint(y.cc, c("x","m"), 0.95)[1,1] # theta1_L
  table.CC[i,3] <- confint(y.cc, c("x","m"), 0.95)[1,2]  # theta1_R

  table.CC[i,4] <- coef(summary(y.cc))[3,1] # theta2
  table.CC[i,5] <- confint(y.cc, c("x","m"), 0.95)[2,1] # theta2_L
  table.CC[i,6] <- confint(y.cc, c("x","m"), 0.95)[2,2] # theta2_R

  table.CC[i,7] <- coef(summary(m.cc))[2,1] # beta
  table.CC[i,8] <- confint(m.cc, "x", 0.95)[1,1] # beta_L
  table.CC[i,9] <- confint(m.cc, "x", 0.95)[1,2] # beta_R

  table.CC[i,10] <- table.CC[i,4] * table.CC[i,7] # IE
  table.CC[i,11] <- table.CC[i,1] + table.CC[i,10] # TE
  table.CC[i,12] <- table.CC[i,10]/table.CC[i,11] # PM

  # fill the med table
  med.table.CC[i,"MIE"] <- med$d0
  med.table.CC[i,"MIE_L"] <- med$d0.ci[[1]]
  med.table.CC[i,"MIE_R"] <- med$d0.ci[[2]]

  med.table.CC[i,"MDE"] <- med$z0
  med.table.CC[i,"MDE_L"] <- med$z0.ci[[1]]
  med.table.CC[i,"MDE_R"] <- med$z0.ci[[2]]

  med.table.CC[i,"MTE"] <- med$tau.coef
  med.table.CC[i,"MTE_L"] <- med$tau.ci[[1]]
  med.table.CC[i,"MTE_R"] <- med$tau.ci[[2]]

  med.table.CC[i,"MPM"] <-  med$n0
  med.table.CC[i,"MPM_L"] <- med$n0.ci[[1]]
  med.table.CC[i,"MPM_R"] <- med$n0.ci[[2]]
}
round(colMeans(table.CC[,c(1,4,7,10:12)]),3)
round(colMeans(med.table.CC[,c(1,4,7,10)]),3)
```

Bias - CC

```
CC.bias <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 6)),
                   c("theta1","theta2", "beta","IE","TE","PM"))
med.CC.bias <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 4)),
                          c("MIE","MDE","MTE","MPM"))

CC.bias[1] <- bias_calculator(table.CC[,1], 100, nsims)
CC.bias[2] <- bias_calculator(table.CC[,4], -50, nsims)
CC.bias[3] <- bias_calculator(table.CC[,7], -0.2, nsims)
CC.bias[4] <- bias_calculator(table.CC[,10], 10, nsims)
CC.bias[5] <- bias_calculator(table.CC[,11], 110, nsims)
CC.bias[6] <- bias_calculator(table.CC[,12], 0.09, nsims)

med.CC.bias[1] <- bias_calculator(med.table.CC[,1], 10, nsims)
med.CC.bias[2] <- bias_calculator(med.table.CC[,4], 100, nsims)
med.CC.bias[3] <- bias_calculator(med.table.CC[,7], 110, nsims)
med.CC.bias[4] <- bias_calculator(med.table.CC[,10], 0.09, nsims)

round(CC.bias,2)
round(med.CC.bias,2) # same
bias.table[,2] <- t(round(CC.bias,2))
bias.table
```

Coverage - CC

```
CC.coverage <- table.CC %>%
  mutate(theta1_coverage = ifelse(theta1_L <= 100 & theta1_R >= 100,
                              T,
                              F)) %>%
  mutate(theta2_coverage = ifelse(theta2_L <= -50 & theta2_R >= -50,
                              T,
                              F)) %>%
  mutate(beta_coverage = ifelse(beta_L <= -0.2  & beta_R >= -0.2,
                              T,
                              F))
med.CC.coverage <- med.table.CC %>%
  mutate(MIE_coverage = ifelse(MIE_L <= 10 & MIE_R >= 10,
                              T,
                              F)) %>%
  mutate(MDE_coverage = ifelse(MDE_L <= 100 & MDE_R >= 100,
                              T,
                              F)) %>%
  mutate(MTE_coverage = ifelse(MTE_L <= 110  & MTE_R >= 110,
                              T,
                              F)) %>%
  mutate(MPM_coverage = ifelse(MPM_L <= 0.09  & MPM_R >= 0.09,
                              T,
                              F))

cc.coverage.table <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 7
)), c("theta1", "theta2","beta","MIE","MDE","MTE","MPM"))

  cc.coverage.table[,1] <- cov_calculator(flags = CC.coverage$theta1_coverage
, nsims = nsims)
  cc.coverage.table[,2] <- cov_calculator(flags = CC.coverage$theta2_coverage
, nsims = nsims)
  cc.coverage.table[,3] <- cov_calculator(flags = CC.coverage$beta_coverage,
nsims = nsims)

  cc.coverage.table[,4] <- cov_calculator(flags = med.CC.coverage$MIE_coverag
e, nsims = nsims)
  cc.coverage.table[,5] <- cov_calculator(flags = med.CC.coverage$MDE_coverag
e, nsims = nsims)
  cc.coverage.table[,6] <- cov_calculator(flags = med.CC.coverage$MTE_coverag
e, nsims = nsims)
  cc.coverage.table[,7] <- cov_calculator(flags = med.CC.coverage$MPM_coverag
e, nsims = nsims)

coverage.table[,2] <- t(cc.coverage.table)
coverage.table
```

Interval length - CC

```
length.CC <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 7)),
                        c("theta1","theta2","beta","MIE","MDE","MTE","MPM"))
length.CC[,1] <- abs(table.CC[,3] - table.CC[,2])
length.CC[,2] <- abs(table.CC[,6] - table.CC[,5])
length.CC[,3] <- abs(table.CC[,9] - table.CC[,8])

length.CC[,4] <- abs(med.table.CC[,3] - med.table.CC[,2])
length.CC[,5] <- abs(med.table.CC[,6] - med.table.CC[,5])
length.CC[,6] <- abs(med.table.CC[,9] - med.table.CC[,8])
length.CC[,7] <- abs(med.table.CC[,12] - med.table.CC[,11])

length.CC.table<- round(colMeans(length.CC),3)
length.table[,2] <- length.CC.table
length.table
```

3. MICE(ML)

```
table.miceML <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 18
)),
                         c("theta1","theta1_L","theta1_R",
                           "theta2", "theta2_L","theta2_R",
                           "beta","beta_L","beta_R",
                           "IE","IE_L","IE_R",
                           "TE", "TE_L","TE_R",
                           "PM","PM_L","PM_R"))

simulations.theta1 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations) #10
00 rows*10 cols
simulations.theta2 <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)
simulations.beta <-  matrix(NA, nrow=n.simulations, ncol=n.imputations)

for (i in 1:nsims) {
  #########################################
  c1 <- rnorm(n, 25, 3)
  c2 <- rbinom(n, 1, 0.5)
  c3 <- rbinom(n, 1, 0.1)
  x <- rbinom(n, 1, 0.5)
  m <- rnorm(n, mean=-2.5 + -0.2 * x + 0.1 * c3, sd=0.8)
  y <- rnorm(n, 3500 + 100 * x + -50*m + 10*c1 + -100*c2 + -200*c3, 300)
  gold.data <- data.frame(y, x, m, c1, c2, c3)

   # make some missingness
  mypattern <-  matrix(c(1,1,0,1,1,1, # only m missing
                         1,1,1,0,1,1, # only c1 missing (BMI)
                         1,1,1,1,1,0, # only c3 missing (smoke)
                         1,1,0,1,1,0), 4,6, byrow = T) # m and c3 missing
  myfreq <- c(0.48, 0.2, 0.2, 0.12) # for each pattern
  myweights <- matrix(c(0, 1, 0, 0.1, 0, 0.3,
                        # missingness of m is correlated to x, c1, c3
                        # x is weighted 10 times as heavy as c1, etc.
                        0, 1, 0, 0, 0, 0,
                        # missingness of BMI is cor to x
                        0, 1, 0, 0.5, 0, 0,
                        # missingness of smoke is cor to x, BMI
                        0, 1, 0, 0, 0.4, 0), 4,6,byrow = T)
                        # missingness of both m & smoke are cor to x, BMI

  miss <- ampute(gold.data,freq = myfreq , patterns = mypattern, weights = my
weights, prop = miss.prop, mech = "MAR")

  observed.data <- miss$amp
  # MICE
  mi.data <- mice(observed.data, m = n.imputations, printFlag = F)
  y.mice <- with(mi.data, lm(y ~ x + m + c1 + c2 + c3))
  m.mice <- with(mi.data, lm(m ~ x + c1 + c2 + c3))
   ############ so far the same ############

  # theta1
  for (a in 1:n.imputations){ # each of the 10 imputed ds
    theta1.hat <- coef(summary(y.mice$analyses[[a]]))[2,1]
    theta1.hat.se <- coef(summary(y.mice$analyses[[a]]))[2,2]
    simulations.theta1[,a] <- rnorm(n.simulations, theta1.hat, theta1.hat.se)
  }

    table.miceML[i,1]<- quantile(simulations.theta1, probs=0.5)
    table.miceML[i,2] <- quantile(simulations.theta1, probs=0.025) # theta1_L
    table.miceML[i,3] <- quantile(simulations.theta1, probs=0.975) # theta1_R

  # theta2
  for (b in 1:n.imputations){
    theta2.hat <- coef(summary(y.mice$analyses[[b]]))[3,1]
    theta2.hat.se <- coef(summary(y.mice$analyses[[b]]))[3,2]
    simulations.theta2[,b] <- rnorm(n.simulations, theta2.hat, theta2.hat.se)
  }

    table.miceML[i,4] <- quantile(simulations.theta2, probs=0.5)
    table.miceML[i,5] <- quantile(simulations.theta2, probs=0.025) # theta2_L
    table.miceML[i,6] <- quantile(simulations.theta2, probs=0.975) # theta2_R

  # beta
  for (c in 1:n.imputations){
    beta.hat <- coef(summary(m.mice$analyses[[c]]))[2,1]
    beta.hat.se <- coef(summary(m.mice$analyses[[c]]))[2,2]
    simulations.beta[,c] <- rnorm(n.simulations, beta.hat, beta.hat.se)
  }
    table.miceML[i,7] <- quantile(simulations.beta, probs=0.5)
    table.miceML[i,8] <- quantile(simulations.beta, probs=0.025) # beta_L
    table.miceML[i,9] <- quantile(simulations.beta, probs=0.975) # beta_R

  # IE
  table.miceML[i,10] <- quantile(simulations.theta2*simulations.beta, probs=0
.5)
  table.miceML[i,11] <- quantile(simulations.theta2*simulations.beta, probs=0
.025)  #IE_L
  table.miceML[i,12] <- quantile(simulations.theta2*simulations.beta, probs=0
.975)  #IE_R
  # TE
  table.miceML[i,13] <- quantile(simulations.theta1 +
                              simulations.theta2*simulations.beta,probs=0.5)
  table.miceML[i,14] <- quantile(simulations.theta1 +
                              simulations.theta2*simulations.beta,probs=0.025
) # TE_L
  table.miceML[i,15] <- quantile(simulations.theta1 +
                              simulations.theta2*simulations.beta, probs=0.97
5)# TE_R
  # PM
  table.miceML[i,16] <- quantile(
    (simulations.theta2*simulations.beta)/
    (simulations.theta1 +simulations.theta2*simulations.beta), probs=0.5)
  table.miceML[i,17] <-quantile(
    (simulations.theta2*simulations.beta)/
```

```
      (simulations.theta1 + simulations.theta2*simulations.beta), probs=0.025
) # PM_L
  table.miceML[i,18] <-quantile(
    (simulations.theta2*simulations.beta)/
      (simulations.theta1 + simulations.theta2*simulations.beta), probs=0.975
) # PM_R
}

colMeans(table.miceML[,c(1,4,7,10,13,16)])
```

Bias - MICE(ML)

```
miceML.bias <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 6)),
                    c("theta1","theta2", "beta","IE","TE","PM"))

  miceML.bias[1] <- bias_calculator(table.miceML[,1], 100, nsims)
  miceML.bias[2] <- bias_calculator(table.miceML[,4], -50, nsims)
  miceML.bias[3] <- bias_calculator(table.miceML[,7], -0.2, nsims)
  miceML.bias[4] <- bias_calculator(table.miceML[,10], 10, nsims)
  miceML.bias[5] <- bias_calculator(table.miceML[,13], 110, nsims)
  miceML.bias[6] <- bias_calculator(table.miceML[,16], 0.09, nsims)


bias.table[,3]<- t(round(miceML.bias,2))
bias.table
```

Coverage - MICE(ML)

```
coverage.miceML <- table.miceML   %>%
  mutate(theta1_coverage = ifelse(theta1_L <= 100 & theta1_R >= 100,
                              T,
                              F)) %>%
  mutate(theta2_coverage = ifelse(theta2_L <= -50 & theta2_R >= -50,
                              T,
                              F)) %>%
  mutate(beta_coverage = ifelse(beta_L <= -0.2  & beta_R >= -0.2,
                              T,
                              F)) %>%
  mutate(IE_coverage = ifelse(IE_L <= 10 & IE_R >= 10,
                              T,
                              F)) %>%
  mutate(TE_coverage = ifelse(TE_L <= 110 & TE_R >= 110,
                              T,
                              F)) %>%
  mutate(PM_coverage = ifelse(PM_L <= 0.09  & PM_R >= 0.09,
                              T,
                              F))
```

```
miceML.cov.table <- setNames(data.frame(matrix(data = NA, nrow = 1, ncol = 7)
), c("theta1", "theta2","beta","IE","DE","TE","PM"))

  miceML.cov.table[,1] <- cov_calculator(flags = coverage.miceML$theta1_cover
age, nsims = nsims)
  miceML.cov.table[,2] <- cov_calculator(flags = coverage.miceML$theta2_cover
age, nsims = nsims)
  miceML.cov.table[,3] <- cov_calculator(flags = coverage.miceML$beta_coverag
e, nsims = nsims)
  miceML.cov.table[,4] <- cov_calculator(flags = coverage.miceML$IE_coverage,
 nsims = nsims)
  miceML.cov.table[,5] <- cov_calculator(flags = coverage.miceML$theta1_cover
age, nsims = nsims)
  miceML.cov.table[,6] <- cov_calculator(flags = coverage.miceML$TE_coverage,
 nsims = nsims)
  miceML.cov.table[,7] <- cov_calculator(flags = coverage.miceML$PM_coverage,
 nsims = nsims)

miceML.cov.table

coverage.table[,3]<- t(round(miceML.cov.table,3))
coverage.table
```

Interval length

```
length.miceML <- setNames(data.frame(matrix(data = NA, nrow = nsims, ncol = 7
)),
                        c("theta1","theta2","beta","IE","TE","PM"))
length.miceML[,1] <- abs(table.miceML[,3] - table.miceML[,2])
length.miceML[,2] <- abs(table.miceML[,6] - table.miceML[,5])
length.miceML[,3] <- abs(table.miceML[,9] - table.miceML[,8])
length.miceML[,4] <- abs(table.miceML[,12] - table.miceML[,11])
length.miceML[,5] <- abs(table.miceML[,3] - table.miceML[,2])
length.miceML[,6] <- abs(table.miceML[,15] - table.miceML[,14])
length.miceML[,7] <- abs(table.miceML[,18] - table.miceML[,17])

length.miceML.table <- round(colMeans(length.miceML),2)
length.table[,3]<- length.miceML.table
length.table
```

Results

```
bias.table
coverage.table
length.table
```