Statistical Inference with Auxiliary Information under Block-Structured Missing Data

Master's thesis in Engineering Mathematics and Computational Science

Linnéa Holmberg

DEPARTMENT OF MATHEMATICAL SCIENCES

Chalmers University of Technology
Gothenburg, Sweden 2025
www.chalmers.se

MASTER'S THESIS 2025

Statistical Inference with Auxiliary Information under Block-Structured Missing Data

Linnéa Holmberg

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025

Statistical Inference with Auxiliary Information under Block-Structured Missing Data
Linnéa Holmberg

Supervisor: Henrik Imberg, Department of Molecular and Clinical Medicine, University of Gothenburg
Examiner: Rebecka Jörnsten, Department of Mathematical Sciences, University of Gothenburg

Master's Thesis 2025
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Statistical Inference with Auxiliary Information under Block-Structured Missing Data
Linnéa Holmberg
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

In medical research, a common challenge is missing data. Missing data can lead to biased findings and loss of precision if not handled appropriately. Common methods of handling missing data are complete-case analysis (CCA), multiple imputation (MI), or inverse probability weighting (IPW), but these methods have drawbacks. This thesis aims to compare these methods to the method augmented inverse probability of complete-case weighting (AIPCCW), that is less established but with certain desirable theoretical properties. AIPCCW is an extension of inverse probability of complete-case weighting (IPCCW), and utilises information from both participants with fully observed data and participants with partly observed data. AIPCCW utilises two models, one for the outcome and one for missingness, where only one model is required to be correctly specified for AIPCCW to achieve unbiased inference.

This thesis implement and compare AIPCCW, CCA, MI, and IPCCW in different scenarios through a simulation study and a application study on real-world data. The scenarios cover unique combinations of sample size, proportion of missing data, levels of correlation between variables with missing values and with an auxiliary variable, and different missingness mechanisms.

In our experiments, the AIPCCW method demonstrate a performance in bias and eRMSE statistically significantly better than CCA, MI, and IPCCW, in certain scenarios, especially in simulated scenarios with a large proportion of missing data. AIPCCW is found to significantly improve in scenarios with a higher correlation between the variable with missing values and the auxiliary variable. On the other hand, the performance of AIPCCW is found to not outperform CCA, MI, and IPCCW in a majority of the scenarios that were implemented. AIPCCW performed comparable to CCA and IPCCW on real-world data in this study, but AIPCCW could potentially perform better on real-world data if a stronger correlation between the variable with missing data and the auxiliary variable existed. Owing to these results, the evaluation is inconclusive to whether AIPCCW is significantly better than CCA, MI, and IPCCW. This thesis concludes that AIPCCW is a stable method, but does not necessarily recommend it over more common methods. However, further research is needed.

Keywords: Augmented inverse probability weighting, block-structured missing data, coarsened data, auxiliary information, missing data, statistical inference, doubly robust methods.

Acknowledgements

I would first and foremost like to extend my great gratitude to my supervisor Henrik Imberg at Statistiska Konsultgruppen Sweden AB, who has shared invaluable insights, thoughts and knowledge, in addition to understanding and patience. I also want to express my sincere appreciation to my examiner Rebecka Jörnsten at Department of Mathematical Sciences, University of Gothenburg, for her helpful feedback and guidance throughout this work. I want to further express how grateful I am for Vilhelm Larsson who has supported me during this process, through both cherished friendship and valuable discussions. I would also like to thank everyone who has dedicated their time and energy to reading, discussing, and providing feedback on this thesis, or supported me in any other way. Completing it would not have been possible without the warm support I have received from dear friends and family.

Linnéa Holmberg, Gothenburg, November 2025

List of Acronyms

Acronyms used in this thesis are listed below in alphabetical order.

AIPCCW	Augmented inverse probability of complete-case weighting
AIPW	Augmented inverse probability weighting
CCA	Complete case analysis
CSF	Cerebrospinal fluid
eRMSE	Estimation root mean squared error
GLM	Generalised linear model
HIV	Human immunodeficiency virus
IL	Interleukin
IPCCW	Inverse probability of complete-case weighting
IPW	Inverse probability weighting
MAR	Missing at random
MARB	Mean absolute relative bias
MCAR	Missing completely at random
MI	Multiple imputation
MLE	Maximum likelihood estimation
MNAR	Missing not at random
PCOS	Polycystic ovary syndrome
PDF	Probability density function
RMSE	Root mean squared error
PROVE	Pre-eclampsia obstetric adverse events
TNF-α	Tumour necrosis factor alpha

Nomenclature

The following nomenclature defines the indices, parameters, and variables used throughout this thesis.

Indices

i, j	Indices for observations and variables
m	Index for simulation runs

Parameters

$\beta_{X,Y}$	Regression coefficient of Y on X
$\hat{\beta}_{X,Y}$	Estimated parameter value of regression parameter $\beta_{X,Y}$.
$\rho_{X,Y}$	Correlation coefficient between X and Y
α_0, α_1, α_2	regression parameters
b, $b1$	regression parameters
θ	True parameter value
$\hat{\theta}$	Estimated parameter value
σ	Standard deviation of the error term

Variables

R_i	Binary indicator variable equal to 1 if observation i is complete, and 0 otherwise
Y, y_i	Response variable Y, and its observed value y_i for individual i
X, x_i	Explanatory variable X, and its observed value x_i for individual i
Z, z_i	Auxiliary variable Z, and its observed value z_i for individual i
e, e_i	Propensity score, and propensity score assigned to observation i

w, w_i	Weight variable, and weight assigned to observation i
ϵ	Error term

Contents

List of Figures

List of Tables

1

Introduction

Medical research and clinical studies are ongoing around the world to cure diseases, discover medical treatments, and to better understand health related problems. This is important for the human population at large, as all medical treatments, vaccines, and medicine, comes from discoveries made in medical research. Vaccination alone have saved approximately 154 million lives globally [1], and in extension, medical research is life saving and can be argued to be crucial for the human population to thrive.

However, medical research frequently encounters challenges which can lead to skewed results and inaccurate conclusions, if not handled appropriately [2]. This is due to missing data. In some cases, missing data is deliberately introduced to reduce ethical concerns (i.e if continued testing are not medically motivated) or financial costs, following a design known as a planned missing data design [3]. These designs often result in block-structured missing data, where specific subsets of participants undergo different levels of data collection.

Planned missing data designs frequently occur in medical research, meanwhile they also lead to challenges as how to handle incomplete observations. This can be handled through deletion of participants with missing values, yet such deletion can lead to inefficiency and bias in the analysis [2]. An alternative approach, MI, aims to re-establish the existing relationship between variables as if no data were missing. However, MI relies on strong assumptions about how data are missing, which makes it difficult to justify when a large proportion of data in key variables of interest are missing [4].

Among the promising methods for handling such missing data is augmented inverse probability of complete-case weighting [5]. AIPCCW is a doubly robust method that combines inverse probability weighting with regression-based adjustment, and has theoretical properties that suggests that it provides more efficient and less biased estimates when handling missing data than IPCCW [6]. Although it offers theoretical benefits, AIPCCW has rarely been used in medical research, and its performance in practical studies remains mostly unexplored.

This thesis present the implementation, and evaluation of AIPCCW as a method for handling missing data in planned missing data designs, with a specific focus on improving inference in medical research. The methodology is demonstrated through simulations and applied to real-world data from pre-eclampsia research. It is also compared to traditional methods of handling missing data used as a benchmark.

1.1 Aim

The aim of this project is to implement and evaluate AIPCCW for handling missing data in planned missing data designs, with a particular focus on block-structured missing data. By assessing the performance of AIPCCW relative to traditional methods such as CCA, MI, and IPCCW, this project aims to compare statistical inference techniques for missing data problems in medical research.

1.2 Specific research questions

The following research questions will be addressed in this project:

- How does the sample size and percentage of missing data affect AIPCCW in comparison to CCA, MI and IPCCW?
- How does the correlation between auxiliary variables and variables with missing data influence AIPCCW's performance?
- How does AIPCCW perform compared to CCA, MI, and IPCCW under different missing data mechanisms?
- Can AIPCCW improve estimation and inference in real-world applications in medical research with block-structured missing data?

1.3 Limitations

This project focuses on developing and evaluating statistical inference methods for handling missing data in measurement-constrained and planned missing data designs, using auxiliary information to improve estimation. The study primarily examines AIPCCW in comparison to traditional approaches, including CCA, MI, and IPCCW. Other techniques, such as Bayesian methods and machine learning-based techniques, fall outside the project's scope.

Additionally, only missing completely at random (MCAR) and missing at random (MAR) scenarios will be considered, where missing values are assumed to be unrelated to unobserved values or explained by fully observed data. Missing not at random (MNAR) mechanisms, where missing values depends on unobserved values, require more complex modelling and are beyond the scope of this project.

1.4 Outline

This thesis is divided into 5 chapters. Chapter 2 provides the theoretical background to missing data mechanisms and several methods of handling missing data. Chapter 3 presents the methods and results of a simulation study on generated data with missing values, to evaluate the performance of the four methods. Chapter 4 similarly presents the method and results of an application study where the four methods are compared to each other, performed on real-world data. Finally, Chapter 5 conclude this thesis through discussing findings presented in Chapter 3 and Chapter 4. References and Appendix are provided at the end of the thesis.

2
Theoretical Background

This chapter opens with a motivating example of missing data, followed by a theoretical background outlining the main missing data mechanisms and the commonly applied methods used to address them. The section provides an overview of the field, serving as a foundation for the methods and applications studied later in this thesis. Further information may be found in e.g. Little & Rubin (2019), van Buuren (2018), and Tsiatis (2006) [7].

2.1 Motivating example

Medical research aims to improve human life, and create opportunities to live a healthier and longer life. To be able to do so, the right conclusion must be feasible, which is only the case if true and accurate data is available. That is, large enough of data must be gathered to be able to answer the question at hand. One interesting and important question is 'How can maternal mortality be lowered?'. This question can be addressed in several ways, however, among the main causes of maternal mortality is pre-eclampsia and is the cause of 47 000 deaths each year globally [8]. Cerebral complications in pre-eclampsia, including eclampsia, cerebral edema, injured blood-brain barrier, and stroke, are the most common causes of maternal mortality [9, 10]. However, pre-eclampsia is a pregnancy disorder that affects multiple organ systems, including brain, liver, kidney, lungs and placenta. If the placenta is affected, then the disorder can cause fetal growth restriction and stillbirth. Furthermore, pre-eclampsia increases the risk of stroke and seizure disorders later in life [11]. Yet no cure, or causes of pre-eclampsia has been discovered.

To better understand these complications, researchers increasingly rely on extensive biomarker and clinical data to explore the underlying mechanisms of the disease. Such studies are inherently complex, as laboratory analyses are resource-intensive and not all tests can be performed on every participant. Consequently, missing data often arise due to financial costs. Effectively addressing such missingness is therefore essential for drawing accurate and reliable conclusions.

2.2 Missing data mechanisms

Missing data can occur both by design and randomly. In the following section different types of missing data mechanisms are introduced. In Figure 2.1 a visualisation of the difference between data missing completely at random and data missing at random conditionally on the variable X on the x-axis is displayed. Note that the red dots illustrating missing data are evenly distributed in the left figure when data are missing completely at

random. Meanwhile the amount of missing data (red dots) in the right figure increases as the value of X increases, i.e when data are MAR given X. The rest of this section is dedicated to the theoretical background of the missing data mechanisms MCAR, MAR, and MNAR.

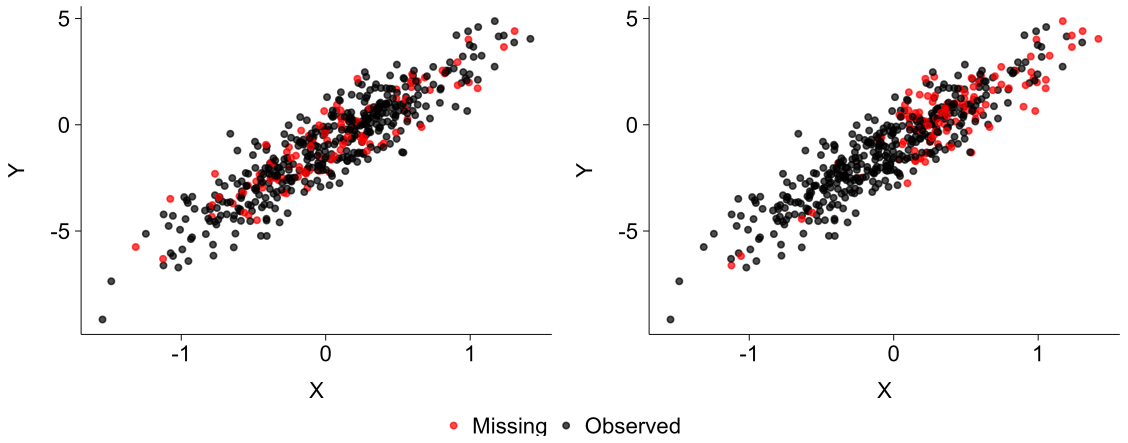

Figure 2.1: Illustration of the missing data mechanisms, with MCAR shown on the left and MAR conditionally on the variable X on the x-axis shown on the right, based on fictitious data. Black dots correspond to observed data, and red dots illustrate unobserved data.

Let R_i be the indicator for individual i being fully observed, i.e having no missing values, defined as

$$R_i = \begin{cases} 1 & \text{If individual } i \text{ is fully observed.} \\ 0 & \text{If individual } i \text{ is not fully observed.} \end{cases}$$

If no data are missing for any participant, that is, $R_i = 1$ for all i, the data is referred to as *complete data* [7]. Let Y denote a continuous response variable of interest (e.g., weight, maximum heart rate, or blood pressure), and let Y_i denote the value of the ith observation of Y.

Consider the collection

$$(R_i, Y_i),\ i = 1, ..., n,$$

where n is the number of participants in the data. Further, we denote the conditional density of R_i given Y_i, noted as $R_i|Y_i$, is defined as

$$P(R_i = 1 \mid Y_i = y_i) =: \pi(y_i),$$

if we let r_i and y_i be the observed values of the random variables Y_i and R_i for observation i. Due to this the joint density of (R_i, Y_i) is

$$p_{R,Y}(r_i, y_i) = p_{R|Y}(r_i|y_i)p_Y(y_i) = \{\pi(y_i)\}^{r_i}\{1 - \pi(y_i)\}^{1-r_i}p_Y(y_i).$$

Furthermore, if R and Y are independent, i.e $R \perp\!\!\!\perp Y$, then missing data do not depend on Y_i. This means that

$$P(R = 1|Y) = \pi(Y) = \pi$$

and data is then said to be MCAR, as the probability π does not depend on any variable. Consequently, the probability of having missing (or observed) values in an observation is completely unrelated to the response variable Y_i, $R \perp\!\!\!\perp Y$, or any otherwise observed or unobserved data.

If missingness is not missing completely at random, then the resulting data and the following analysis can become biased if not handled properly. Let Z denote a fully observed covariate for all i, and assume that missingness in Y depend on Z. Then consider the collection of variables $(R_i, Y_i, Z_i), i = 1, .., n$, where $R \perp\!\!\!\perp Y|Z$. This can be expressed as

$$P(R_i = 1|Y_i, Z_i) = P(R_i = 1|Z_i) = \pi(Z_i), \tag{2.1}$$

and the data are said to be MAR (conditionally on Z, alternatively, given Z), where the probability of participant i having partly observed data is a function of Z. Note that this function of Z is stated in Equation (2.1) as $\pi(Z_i)$. If such a covariate Z exists in the data and are not of primary interest, then such variable is often referred to as an *auxiliary variable* [7]. On the other hand, if the probability of missing data depend on Z but Z is unobserved, then the data is MNAR. If the probability of missingness depend on observed data, i.e MAR, then the relationship for missingness could be modelled as a function of the dependent variables. However, when the probability of missingness depend on unobserved data, i.e MNAR, this can not be modelled as easily and strong assumptions need to be made [7]. The missing data mechanism is important, as to handle missing data appropriately depends on which missing data mechanism that it is. When a method has been chosen, a statistical modelling of interest can be applied. Let X denote an explanatory variable of interest (e.g., height, age, or BMI). A model of interest can then be $Y|X$, and a simple linear regression can be applied, modelled as

$$Y = b_0 + b_1 X,$$

where b_0 is the intercept and b_1 is the regression coefficient associated with X, representing its effect on Y. Let $\theta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$, and the estimated value of the regression parameters be denoted as $\hat{\theta} = \begin{bmatrix} \hat{b_0} \\ \hat{b_1} \end{bmatrix}$.

2.3 Methods for handling missing data

Several number of methods can be used to handle missing data. One of the simplest methods commonly used according to Ross et al [12], is to remove observations containing missing values in any variable. That is, all data (R_i, Y_i, X_i, Z_i) for all i such that $R_i = 1$ are included in the analysis. This method of excluding observations with missing data is often referred to as complete-case analysis. The method introduces no bias when data are MCAR, but may produce biased inference if the data are MAR. If the data are MCAR, then the same relation between variables is captured by fewer observations and the estimators may still behave similarly due to independence between R and other variables [13]. Nonetheless, with fewer observation the analysis would have less statistical power and thus a higher uncertainty [12]. This generally means that CCA have a loss of precision, even if using CCA when data are MCAR as it would not inflict bias as it would

when performed on data that are MAR. However, it is not possible to prove if data are MAR or MCAR, and an assumption of MCAR is restrictive in addition to the possible introduction of bias into the analysis if the assumption is incorrect [12].

Aside from handling missing data through CCA, Tsiatis [7] denote three main approaches to handle missing data to estimate $\hat{\theta}$, consisting of likelihood estimation, imputation methods and IPCCW. The rest of this section is dedicated to explore methods that handle missing data through imputation and IPCCW, in addition to AIPCCW.

2.3.1 Simple imputation methods

To allow observations with missing values to contribute to the statistical analysis, imputation can be used to replace the missing entries with specific values. This can be done once for each missing value, which is called single imputation, or several times, called multiple imputation. Single-value imputation methods, or simple imputation methods, are for example mean imputation, stochastic regression imputation, and last observation carried forward. These methods impute the mean of the variable that contain missing values, a fitted value from a regression model, or impute the last observation of the variable respectively. However, single-value imputation approaches are suboptimal as they can generate biased results and false precision [4].

2.3.2 Multiple imputation

Multiple imputation creates multiple possible values through creating multiple data sets, which is used to perform multiple analyses which are thereafter pooled. This provides a quantification of the uncertainty of the imputation [4].

MI produce numerous possible values through the following two step procedure:

1. Generate alternative values for each missing value multiple times, based on statistical characteristics of the data, resulting in multiple complete data sets.
2. All data sets are analysed separately and thereafter the resulting estimates are pooled.

Further, Li [4] state that MI is often a preferable approach to CCA, as the assumption of MCAR is restrictive meanwhile the uncertainty in MI is quantifiable. The result of analysis using MI can be interpreted as if all participants were fully observed, given that all assumptions of MI are met, i.e. data being MAR and the imputation model being correctly specified [4]. MI attempts to capture the associations present in the full data while accounting for the uncertainty introduced by missing data. MI utilises the conditional distribution of the missing data given the observed data to estimate plausible values and thus being able to retrieve unbiased and consistent results, under the assumption of MI being fulfilled [4]. Yet when only a small number of imputations are performed, MI is affected by stochastic noise [6]. However, this method does not take the probability of the observation being missing in to account, which will be discussed next.

2.3.3 Inverse probability of complete-case weighting

The third approach of handling missing data according to Tsiatis is to take the probability of missing values in to account, which inverse probability of complete-case weighting does [7]. This is possible through the use of propensity scores, which in this case the

propensity of complete-case represent the probability of an individual having no missing values given a set of variables. The fitted probability can be calculated by using a machine learning algorithm or logistic regression.

Let Y_i denote the values of Y, and $\mathbf{X}_i$ denote the values of covariates $\mathbf{X}$, for individual i, in addition to the previously defined variable R_i, i.e the binary indicator assigned to participant i indicating whether they have complete data. We further define propensity scores in this context as $e_i = \Pr(R_i = 1 \mid X_i, Z_i)$. Then a logistic regression model to estimate propensity scores can be defined as

$$\text{logit}(e_i) = \log\left(\frac{e_i}{1 - e_i}\right) = \alpha_0 + \alpha_1 X_i + \alpha_2 Z_i, \tag{2.2}$$

estimating the probability of value Y being observed in participant i, based on information retrieved in X_i and Z_i.

In (2.2) the regression parameters α_0, α_1, and α_2 can be estimated through the use of maximum likelihood estimation (MLE), and is the solution to

$$L(\alpha_0, \alpha_1, \alpha_2) = \prod_{i=1}^{n} e_i^{R_i}(1 - e_i)^{1-R_i}.$$

More generally, we may consider a parameter β of a statistical model f(y|x), describing the relationship between the explanatory variable X and the response variable Y, with corresponding estimator $\hat{\beta}$ defined as the unique solution [6] to the system of equations

$$\sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0}.$$

This covers most methods for estimation in statistics, including, e.g. least squares regression and likelihood-based methods. Utilizing the statistical model f(y|x) and the estimated $\hat{\boldsymbol{\beta}}$, predicted values for each observation can be retrieved, which is the propensity scores e_i assigned to individual i. Weights assigned to each individual i, used in IPCCW, is calculated through [14]

$$w_i = \frac{R_i}{e_i} + \frac{1 - R_i}{1 - e_i} \tag{2.3}$$

where R_i denotes if complete data for individual i were observed, and e_i represents the propensity score. Note that w_i is zero for participants with missing data and equals $\frac{1}{e_i}$, i.e. the inverse propensity of being a complete-case, for individuals with fully observed data. Hence, the analysis can be evaluated on the complete cases. Thus the model only utilise information from individuals that has no missing values. Seaman and White [6] describes how the IPCCW estimator is found as the solution to

$$\sum_{i=1}^{n} R_i w_i \mathbf{U_i}(\boldsymbol{\beta}) = \mathbf{0}. \tag{2.4}$$

In practice, the resulting procedure of implementing IPCCW is presented in Algorithm 1.

IPCCW can be used for several reasons and can reduce bias, adjust for unequal sampling fractions, or deal with missing data [6]. A known drawback of IPCCW is the potential instability of the weights. Individuals with a low propensity score may receive excessively large weights if not properly controlled, which can be mitigated through weight

Algorithm 1 Inverse probability of complete case weighting [5].

1: Fit a logistic regression model for the probability of having complete data, given the remaining factors that has no missing values. If the variable Y contains missing values, then the model of interest would be $R|X, Z$, given that the data contains the three variables Y, X, and Z. Let e_i denote the fitted probability of having complete data for participant i.
2: Fit the desired target statistical model for Y given X, i.e. $Y|X, Z$, on the complete cases, using the observation weights $w_i = 1/e_i$.

trimming or ridge regression. Individuals assigned large weights, can result in IPCCW estimators behaving very badly, owing to (2.3), and can be due to their background characteristics being correlated with the missingness of the response variable, making them very influential [5]. Another reason for individuals receiving large weights, can be due to overfitting of the propensity score model. To improve IPCCW, weight stabilisation can be used, such as trimming of extreme weights or the use of ridge regression when modelling propensity scores, which uses regularisation to handle multicollinearity within the data or possible overfittning of the propensity score model. However, trimming weights can lead to re-introduction of bias which IPCCW aimed to remove. An alternative to weight stabilization for improving IPCCW is to use augmented IPCCW [6].

2.3.4 Augmented inverse probability of complete-case weighting

A extension of IPCCW, or a generalisation of IPCCW, is the doubly robust method augmented inverse probability of complete-case weighting [5]. The difference between IPCCW and AIPCCW is the use of an additional model, which extracts information from partly unobserved participants by regression and fitted values, in addition to the weighted participants information. This results in additional data and information from participants with missing values can thus be utilised, which IPCCW does not include into the analysis. Furthermore, only one of the two models used in AIPCCW need to be correctly specified for AIPCCW to be valid [5], which is why it is referred to as a doubly robust method.

The AIPCCW estimator is calculated through the use of an missingness model and a imputation model, where the first model calculates propensity scores to be used as weights in the second model, similarly to IPCCW. The AIPCCW estimators, i.e β, are found through solving the following equations [6],

$$\sum_{i=1}^{n} R_i w_i \mathbf{U}_i(\boldsymbol{\beta}) + (1 - R_i w_i)\phi_i(\boldsymbol{\beta}) = \mathbf{0}, \tag{2.5}$$

where $\phi_i(\boldsymbol{\beta})$ represents the estimate of the expectation of $\mathbf{U}_i$ on individual i, given observed data. Observe how (2.5) is an extension of (2.4), and the second term corresponds to not fully observed participants and thus utilises information from participants with missing data. The estimators of AIPCCW, given that the modelling of both the missingness model and the imputation model are correct, are at least as precise as the IPCCW estimator [5] in large sample sizes and more efficient than IPCCW [6]. A challenge with AIPCCW according to Seaman and White is the lack of software and how doubly robust methods that is asymptotically equivalent to AIPCCW is a possible solution [6]. In this thesis

a description of such procedures, published by Vansteelandt et al, are implemented and utilised [5]. Seaman and White further states that an analytical calculation of standard errors is challenging, but a possible way of handling this problem is through the use of bootstrap [6].

Furthermore, as the auxiliary variable Z needs to be correlated with the variable or variables that have missing values, in addition to the adjustments needed to model based on which variables have missing values, the procedure of APICCW differs. This limits the generalisability of AIPCCW, as missing in two variables, or in both response variable and explanatory variable, result in a more complex modelling. Vansteelandt et al [5] present procedures for two special cases of handling missing data when the response variable is continuous through the use of AIPCCW, which this thesis focuses on. The first special case is when there are missing values in the response variable, and the procedure of handling this is presented in Algortihm 2. The second special case is when there are missing values in a single explanatory variable, and there exists an additional explanatory variable, which is presented in Algortihm 3.

Algorithm 2 AIPCCW when missing values occur in the outcome variable [5].

Require: Data with missing values only in the response variable.

1: Fit a logistic regression model for the probability of observing $\mathbf{Y_i}$ for participant i given the remaining factors that have no missing values. That is, model a logistic regression $R|X, Z$ where Z is an auxiliary variable correlated with Y_i. Let e_i denote the fitted value, i.e the propensity score, for participant i.
2: Fit a weighted generalized linear regression model to the fully observed participants using weights $1/e_i$. That is, model $Y|X, Z$. Let $f^*(X_i, Z_i)$ denote the predicted value for participant i.
3: Fit a generalized linear model $f^*(X_i, Z_i)|X$ using the full data set. The estimated parameter value assigned to X is the AIPCCW estimator.

Algorithm 3 AIPCCW when missing values occur in a single covariate [5], and the response variable is continuous.

Require: Data with missingness in only one covariate.

1: Fit a logistic regression model for the probability of observing $\mathbf{X_i}$ for participant i given the remaining factors that has no missing values, which need to include at least the response variable, an auxiliary variable correlated to X, and an additional fully observed covariate. Let the additional covariate be denoted by Q. If no additional covariate exists, add an additional variable $Q = \mathbf{1}$. The logistic regression model is then $R|Y, Z, Q$. Let $_i$ denote the fitted value, i.e the propensity score, for participant i.

2: Fit a linear weighted regression model of interest which in our case is $Y|X, Q$, using the observed data with weights $1/e_i$. Denote the estimated parameter value assigned to X to be β_X, and the estimated parameter value assigned to Q to be denoted by β_Q.

3: Estimate $E(X)$, which in our case is modelled through a linear regression model $X|Q, Y, Z$ with no intercept, using weights $1/e$. Denote the predicted values from the given model to be $\hat{X}^1$.

4: Estimate $E(X^2)$, which in our case is modelled through a linear regression model with only an intercept and use $\hat{X}^1$ as an offset term. Denote the predicted values from the given model to be $\hat{X}^2$.

5: Set a tolerance threshold and denote it as *tolerance_threshold*. Set a variable named *iteration* to zero, and decide a maximum number of iterations. Declare a new variable named *maximum_iteration* to the number of maximum iterations. Set a variable named *converged* to be false.

6: **while** (*converged*=FALSE **and** *iteration* <*maximum_iteration*) **do**

7: Fit a new model $Y - \beta_x \hat{X}^1 | Q$ with no intercept from data where $R = 1$. Denote the estimated parameter value assigned to covariate Q to be β_{Qnew}.

8: Fit a new model $Y - \beta_{Qnew} Q\hat{X}^1 / \sqrt{\hat{X}^2} | \sqrt{\hat{X}^2}$ with no intercept. Denote the estimated parameter value of $\sqrt{\hat{X}^2}$ to β_{Xnew}

9: **if** ($|\beta_X - \beta_{Xnew}| <$ *tolerance_threshold* **and** $|\beta_Q - \beta_{Qnew}| <$ *tolerance_threshold*) **then**

10: *converged* $\leftarrow$ TRUE

11: **end if**

12: $\beta_Q \leftarrow \beta_{Qnew}$

13: $\beta_X \leftarrow \beta_{Xnew}$

14: **end while**

15: The resulting estimated parameter value β_X is our AIPCCW estimator.

3

Simulation study

In this chapter, a simulation study evaluating AIPCCW is carried out and the methodology and results are presented. Various scenarios are explored, including different missingness mechanisms and cases where missing data occur in either the response or explanatory variables. Firstly, the objectives for the simulation study are presented in Section 3.1, followed by simulation setup in Section 3.2, formulation of model in Section 3.2.1, simulation settings in Section 3.2.2, and evaluation metrics in Section 3.2.3. Lastly, this chapter presents the results from the simulation study in Section 3.3.

3.1 Introduction and objectives

The objectives of the simulation study were to answer the following research questions introduced in 1:

1. How does sample size and proportion of missing data affect AIPCCW in comparison to CCA, MI, and IPCCW?
2. How does the correlation between the auxiliary variable and the variable with missing data influence AIPCCW's performance?
3. How does AIPCCW perform compared to CCA, MI, and IPCCW under different missing data mechanisms?

3.2 Simulation setup

Twelve different simulation setups were considered, depending on whether missingness occurred in X or Y, which mechanism was assumed (MCAR, MAR given X, or MAR given Z), and whether X was continuous or binary. This resulted in 12 combinations of different simulation setups which the methods were evaluated on, as detailed in Section 3.2.2.

In each simulation scenario, a complete dataset containing 100,000 observations was simulated in order to retrieve the true parameter value describing the linear relationship between X and Y, through linear regression further described in Section 3.2.1. The coefficient of determination, R^2, was similarly retrieved from the modelling of the large dataset. R^2 was calculated with the following formula

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i denotes the observed value for individual i, and $\hat{y}_i$ denotes the fitted value for individual i retrieved from the by the linear regression. The coefficient of determination was

controlled through the variance of the error term, where increasing values of the variance decreased the overall, whereas decreasing values increased it. This was followed by randomly selecting a sample of n observations, and out of those n observations a proportion of observations were randomly chosen to be missing, either according to MCAR, MAR given X, or MAR given Z, in the response variable Y or the explanatory variable X. The response variable Y were set to be missing for a random sample of observations using different approches depending on the missing data mechanism. If data in Y were MCAR, a random selection of observations was set to have missing values in Y. If the data in Y were MAR conditionally dependent on X or Z, an probability weight was assigned to each observation according to its value in X or Z respectively. After each data generating process finished, the different methods were applied on the same data set and evaluation metrics for each method computed.

To answer research question 2 which aimed at understanding how correlation between the auxiliary variable and the variable with missing data affect AIPCCW, the data generating procedure differed depending on whether missingness occurred in Y or X. Thus generating the data set $\{X, Y, Z\}$, and the formulation of the two models also differed. In the following section the formulation of the two models are described.

3.2.1 Formulation of the model

3.2.1.1 Missing data in the response variable

The data were simulated using linear regression and all variables assumed to be normally distributed. We let X be drawn from a normal distribution, with mean $\mu = 0$ and variance $\sigma^2 = 1$. Note that the choice of values of μ and σ are arbitrary, and this choice does not affect the result. The following model was used to generate data with missing values in Y in the following order:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon_s \\ Z &= \beta_{Z,Y} Y + \epsilon_z. \end{aligned} \tag{3.1}$$

where the error term ϵ_s was assumed to be normally distributed , with $\mu_{\epsilon_s} = 0$ and $\sigma^2_{\epsilon_s} = s^2$ to vary during different simulation runs to asses the models at different levels of R^2. However, the error term ϵ_z is assumed to be drawn from a normal distribution, with $\mu_{\epsilon_z} = 0$ and $\sigma^2_{\epsilon_z} = 1$. During the simulation we let $\beta_0 = 0$, and $\beta_1 = 1$. We denote the corresponding estimate of β_1 by $\hat{\beta}$. Additionally, the regression coefficient $\beta_{Z,Y}$ is included to control the correlation between the variables Z and Y. This allows the methods to be evaluated at different correlation levels between the variable with missing values, Y, and the auxiliary variable Z, addressing research question 2. Note that regardless of the specific research questions, an auxiliary variable that is correlated with the variable containing missing values must be present in order for AIPCCW to be applied. The regression coefficient $\beta_{Z,Y}$, the variance in Y, and variance in Z, was used to calculate the correlation between the two variables in the data denoted $\rho_{Z,Y}$, as

$$\rho_{Z,Y} = \beta_{Z,Y} \frac{\sqrt{Var(Z)}}{\sqrt{Var(Y)}}$$

The modelling of the relations used for further statistical analysis is visualised in Figure 3.1 as a directed acyclic graph, with noted correlation.

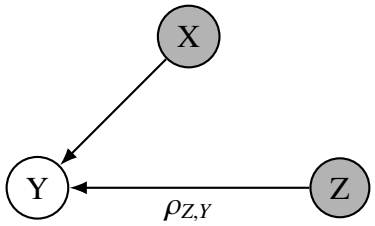

Figure 3.1: Directed acyclic graph visualising the modelling of simulated data when missing values occur in the response variable Y. Grey nodes correspond to fully observed data and white nodes illustrate partly observed data.

3.2.1.2 Missing data in the explanatory variable

Data with missing values in the explanatory variable X, were similarly generated as described in the previous section. The difference is the modelling of data, where we in this scenario instead let X be correlated with Z. The reason for this is that the research questions aims at understanding how AIPCCW and other methods perform when using auxiliary variables correlated to variables with missing data. Thus, let X and Z be drawn from a multivariable normal distribution with correlation $\rho_{Z,X}$, and let Y be generated by the two variables. That is, X and Z generated with the probability density function (PDF)

$$f(x,z) = \frac{1}{2\pi\sigma_X\sigma_Z\sqrt{1-\rho_{X,Z}^2}} exp\Big(-\frac{1}{2[1-\rho^2]}\Big[\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 - 2\rho_{X,Z}\Big(\frac{x-\mu_X}{\sigma_X}\Big)\Big(\frac{z-\mu_Z}{\sigma_Z}\Big) + \Big(\frac{z-\mu_Z}{\sigma_Z}\Big)^2\Big]\Big).$$

If we set $\mu = \begin{bmatrix}\mu_X \\ \mu_Z\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}$ and $\sigma = \begin{bmatrix}\sigma_X \\ \sigma_Z\end{bmatrix} = \begin{bmatrix}1 \\ 1\end{bmatrix}$, this results in the PDF

$$f(x,z) = \frac{1}{2\pi\sqrt{1-\rho_{X,Z}^2}} exp\Big(-\frac{1}{2[1-\rho^2]}\Big[x^2 - 2\rho_{X,Z}xz + z^2\Big]\Big).$$

From X and a fixed constant $Q = 1$, we simulate Y according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Q + \epsilon_s. \tag{3.2}$$

The constant Q is introduced due to the algorithm requires an additional explanatory variable to iterate between in order to retrieve the AIPCCW estimator, and this constant can be viewed as an additional intercept. We similarly to the previous section define ϵ_s as an added normally distributed error term, with $\mu_{\epsilon_s} = 0$ and $\sigma^2_{\epsilon_s} = s$, and let s vary during different simulation runs to assess the models at different levels of R^2. During the simulation we let $\beta_0 = 0$, $\beta_1 = 1$, and $\beta_2 = 1$. We denote the corresponding estimate of β_1 by $\hat{\beta}$. Consequently, the modelling of the relation used for further statistical analysis is visualised in Figure 3.2 as a directed graph, with noted correlation.

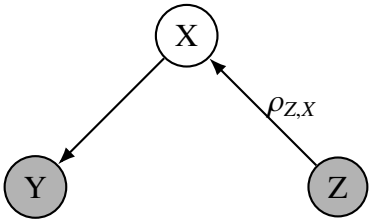

Figure 3.2: Directed acyclic graph visualising the modelling of simulated data when missing values occur in the explanatory variable X. Grey nodes correspond to fully observed data and white nodes illustrate partly observed data.

3.2.2 Simulation settings

To evaluate the performance of AIPCCW in comparison to CCA, MI, and IPCCW in different scenarios, several settings were used. The following settings were tested during the simulation study;

- Sample size $\{40, 80, 150, 300, 450, 600\}$.
- Regression parameter $\beta_{Z,Y}$ in (3.1) to control the related correlation level $\rho_{Z,Y}$, as well as correlation levels $\rho_{Z,X}$ in (3.2) $\{0.2, 0.5, 0.8\}$.
- proportion of missing values $\{0.2, 0.6, 0.85\}$.
- Coefficient of determination R^2 through changing the variance of the error term ϵ_s in (3.1) and (3.2), resulting in $R^2 \in \{0.2, 0.5, 0.8\}$.
- Missing data mechanisms {MCAR, MAR given X, MAR given Z}.
- Variable with missing data {X, Y}.

The above settings were chosen in order to gain insight in the performance of the four methods during a wide range of scenarios. The simulation study evaluated AIPCCW, CCA, MI, and IPCCW for all unique combination of the above settings, and each simulation was repeated 500 times.

3.2.3 Evaluation metrics

The performance of the methods were evaluated by several measures, including the estimation root mean squared error (eRMSE) of the estimator compared to ground truth, defined as

$$\text{eRMSE}(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}_{1m} - \beta_{1m})^2},$$

where M denote the number of repeated simulation runs. The eRMSE was used in order to answer research question regarding the performance of methods in the simulation study. Additionally, an estimation and comparison of bias used to further explore the behaviour of the methods. Bias is the difference between an estimate and the true value, which were calculated through

$$bias = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{1m} - \beta_{1m},$$

where $m = 1, ..., M$ and denotes simulation run number m and $\beta_{1m} \approx 1$. Additionally the mean absolute relative bias (MARB) was studied, which we define as

$$\text{MARB} = \frac{1}{M} \sum_{m=1}^{M} \left| 1 - \frac{\hat{\beta}_{1m}}{\beta_{1m}} \right|,$$

as used by [15].

3.3 Results

This section is divided into two subsections. The first subsection presents results from simulations with missing values in Y, while the second subsection focuses on results from simulations involving missing values in X.

3.3.1 Missing data in the response variable

To study how sample size and the proportion of missing data affect AIPCCW's performance, and thus answer research question 1, we analyse these factors in Figure 3.3. The figure shows how they, together with the correlation between the auxiliary variable Z and the variable with missing values, i.e Y, influence the four methods AIPCCW, CCA, MI, and IPCCW when data are MAR conditional on Z. In Figure 3.3 it can be seen the performance of CCA, MI, and IPCCW are not effected by an increasing correlation between Z and Y, as they do not utilise the information in Z, meanwhile AIPCCW performance of eRMSE decrease with statistical significance when the sample size is large enough. Given a smaller sample size of 40 individuals, no statistically significant decrease in eRMSE for AIPCCW nor statistically significant smaller eRMSE of AIPCCW in comparison to any other method can be seen when the correlation $\rho_{Z,Y}$ increases. The performance of AIPCCW in bias as an evaluation metric compared to CCA, can be found in Table 3.1, where we found a significant difference when data were MAR given Z, that is, the same missing data mechanism that are simulated in Figure 3.3.

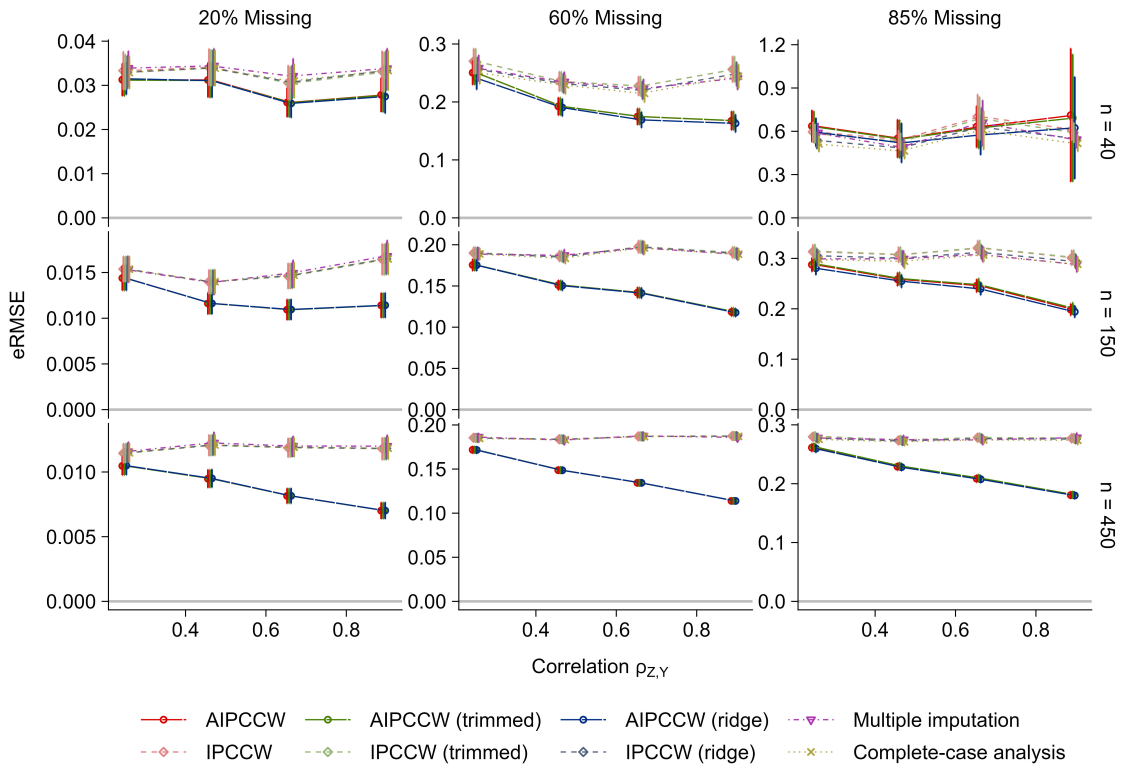

Figure 3.3: Visualisation of estimation RMSE as a function of the correlation $\rho_{Z,Y}$ between the auxiliary and the response variable, comparing different methods of handling missing data at various sample sizes denoted as n, and proportions of data that are MAR conditionally on Z in the response variable.

Table 3.1: Bias of AIPCCW and CCA, and their mean difference, across different sample sizes, proportions of missing data, and missing data mechanisms.

	20% Missing data			85% Missing data		
	AIPCCW Mean (SD)	CCA Mean (SD)	Mean difference est (95% CI) p	AIPCCW Mean (SD)	CCA Mean (SD)	Mean difference est (95% CI) p
n = 40						
MCAR	0.02 (0.40)	0.02 (0.43)	0.00 (-0.01, 0.02)	-0.04 (0.95)	0.02 (1.33)	-0.06 (-0.16, 0.05)
MAR given X	0.00 (0.43)	0.02 (0.48)	-0.02 (-0.04, 0.00)	-0.13 (1.20)	-0.07 (1.70)	-0.06 (-0.19, 0.07)
MAR given Z	-0.14 (0.40)	-0.34 (0.43)	0.20 (0.19, 0.22)***	-0.71 (1.07)	-1.73 (1.16)	1.02 (0.91, 1.13)***
n = 150						
MCAR	−0.00(0.20)	0.00 (0.21)	−0.00(−0.01, 0.00)	−0.01(0.33)	−0.03(0.50)	0.02(−0.02, 0.05)
MAR given X	−0.02(0.20)	−0.02(0.23)	0.00(−0.01, 0.01)	0.01 (0.44)	0.01 (0.54)	−0.00(−0.05, 0.05)
MAR given Z	−0.08(0.21)	−0.27(0.23)	0.19 (0.19, 0.20)***	−0.67(0.34)	−1.57(0.43)	0.90 (0.87, 0.93)***
n = 450						
MCAR	0.00 (0.12)	0.01 (0.13)	−0.00(−0.01, 0.00)	0.00 (0.18)	0.02 (0.28)	−0.02(−0.04, 0.00)
MAR given X	0.00 (0.12)	0.00 (0.13)	0.00(−0.00, 0.01)	−0.00(0.32)	0.01 (0.28)	−0.02(−0.05, 0.01)
MAR given Z	−0.09(0.12)	−0.29(0.13)	0.20 (0.19, 0.20)***	−0.65(0.18)	−1.51(0.22)	0.86 (0.84, 0.88)***

Biases are reported as mean (SD) for each model, based on 500 simulated replicates. Mean differences are presented with corresponding 95% confidence intervals, and p-values were derived from paired t-tests.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
Abbreviations: AIPCCW, augmented inverse probability of complete-case weighting; CCA, complete-case analysis; CI, confidence interval; est, estimate; SD, standard deviation; p, p-value.

Additionally, Figure 3.3 display three different approaches of calculating inverse propensity weights for AIPCCW and IPCCW, that is, without weight stabilisation, with

weight trimming, and with weight stabilisation through ridge regression, which are displayed with the same line type. It is found that no significant difference between the three approaches of handling the weights, as Table A.1 in Appendix presented no statistically significant mean difference in any combination of sample size (40, 80, 150, 450, 600), proportion of missing data (20%, 60%, 85%), or missing data mechanism (MCAR, MAR conditionally on X, MAR conditionally on Z).

Moreover, in Figure 3.4 we visualise the difference in performance under missing data mechanism MCAR, MAR conditionally on X, and MAR conditionally on Z, to address research question 3 - "How does AIPCCW perform compared to CCA, MI, and IPCCW under different missing data mechanisms?".

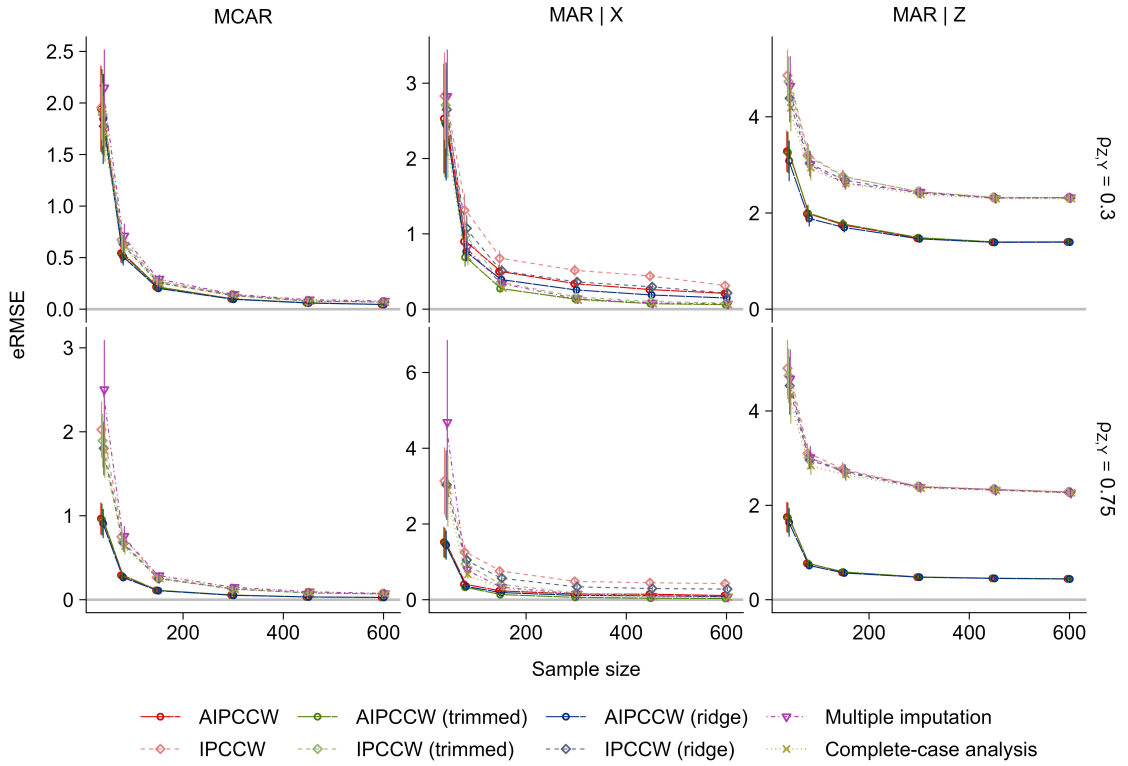

Figure 3.4: Visualisation of estimation RMSE as a function of sample size, comparing different methods of handling 85% missing data for three different missing data mechanisms, i.e MCAR, MAR conditionally on X, and MAR conditionally on Z.

Figure 3.4 display eRMSE as a function of sample size, and we found for all methods a significant decrease in eRMSE given a larger sample size, under all missing data mechanisms. AIPCCW are found to perform significantly better than CCA, MI, and IPCCW under data that are MAR conditionally on Z, as found in previously, in addition to a significantly lower eRMSE when data are MCAR or MAR conditionally on X when the sample size is less than 200 and there is a large correlation level between Z and Y. However, even if Figure 3.4 display a significantly lower eRMSE of AIPCCW than CCA, MI, and IPCCW in some cases during both MCAR or MAR, a gain in performance in bias can only be found when data are MAR conditionally on Z. This is found in Figure 3.5, which visualises the performance of the same factors as in Figure 3.4 but with bias as a

evaluation metric. AIPCCW is found to decrease in eRMSE and bias at a similar rate as CCA, MI, and IPCCW, under all three missing data mechanisms.

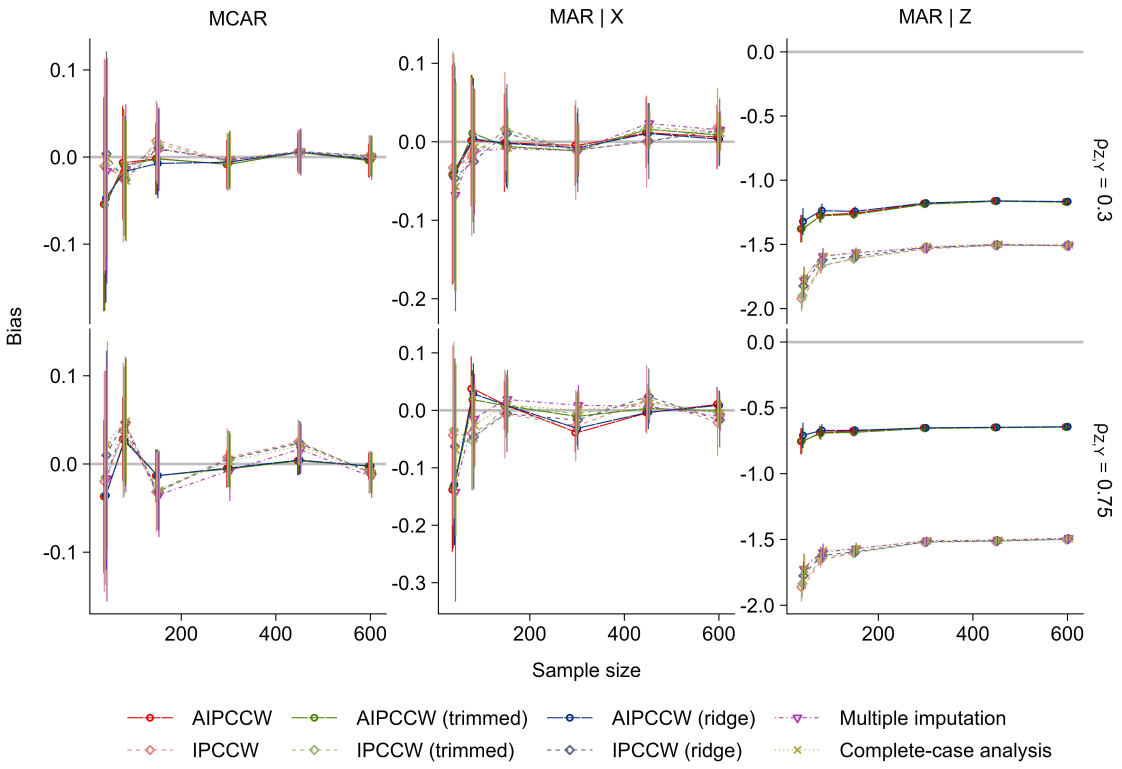

Figure 3.5: Visualisation of bias as a function of sample size, comparing different methods of handling 85% missing data for three different missing data mechanisms, i.e MCAR, MAR conditionally on X, and MAR conditionally on Z.

As displayed in Figure 3.4, a slight decrease of eRMSE for AIPCCW can be found under data that are MAR conditionally on X, comparing $\rho_{Z,Y} = 0.3$, and $\rho_{Z,Y} = 0.75$. However, in Figure 3.6 which displays bias as a function of sample size when data are MAR conditionally on X, it is found that AIPCCW do not significantly improve under either a higher correlation level or different proportion of missing data. Multiple imputation is found to perform badly when the sample size is small and we have 85% missing data. Under a different level R^2, bias are displayed and compared between AIPCCW and MI in Table 3.2, and a similar comparison between AIPCCW and IPCCW is found in Table 3.3.

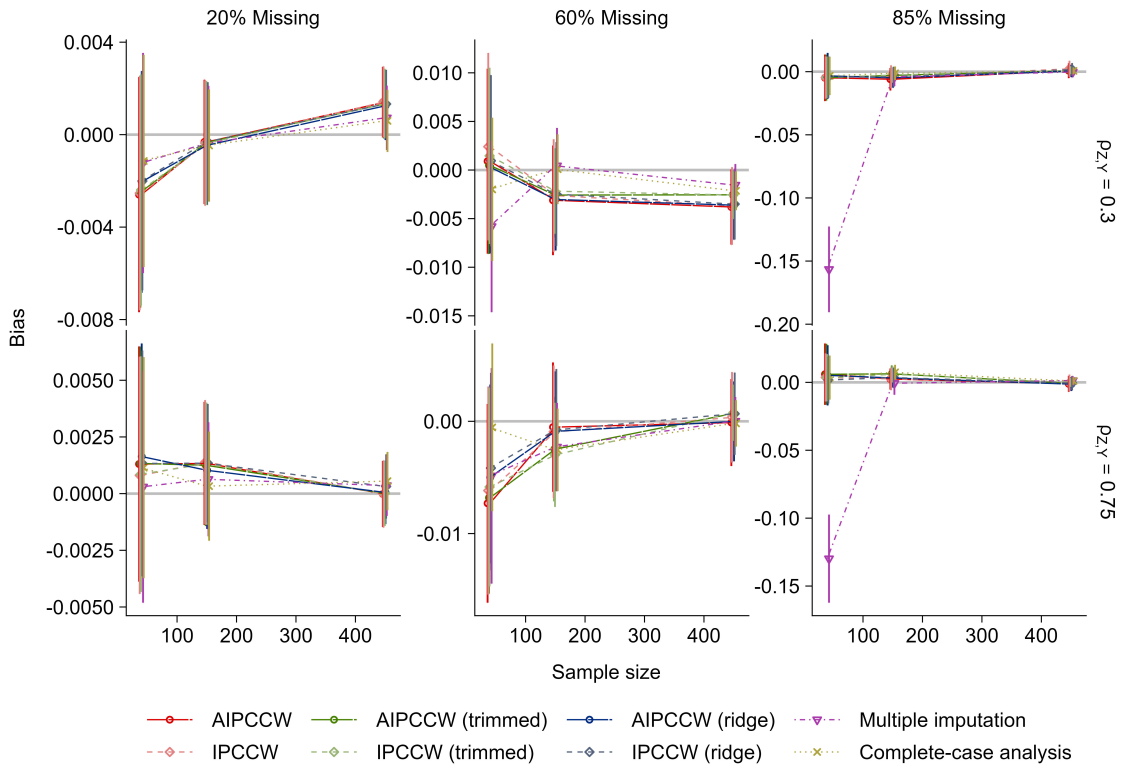

Figure 3.6: Visualisation of bias as a function of sample size, comparing different methods of handling missing data at various levels of correlation $\rho_{Z,Y}$ between the auxiliary variable Z and the response variable Y containing missing values, and various proportions of data that are MAR conditionally on Z.

Table 3.2: Bias of AIPCCW and MI, and their mean difference, across different sample sizes, proportions of missing data, and missing data mechanisms.

	20% Missing data			85% Missing data		
	AIPCCW Mean (SD)	MI Mean (SD)	Mean difference est (95% CI) p	AIPCCW Mean (SD)	MI Mean (SD)	Mean difference est (95% CI) p
n = 40						
MCAR	0.02 (0.40)	0.02 (0.44)	0.01(−0.01, 0.02)	−0.04(0.95)	−0.02(1.58)	−0.02(−0.14, 0.10)
MAR given X	0.00 (0.43)	0.02 (0.49)	−0.02(−0.04, 0.01)	−0.13(1.20)	−0.14(2.16)	0.01(−0.17, 0.19)
MAR given Z	−0.14(0.40)	−0.35(0.44)	0.21 (0.20, 0.23)***	−0.71(1.07)	−1.73(1.31)	1.02 (0.90, 1.13)***
n = 150						
MCAR	−0.00(0.20)	0.00 (0.22)	−0.00(−0.01, 0.01)	−0.01(0.33)	−0.04(0.54)	0.02(−0.02, 0.06)
MAR given X	−0.02(0.20)	−0.02(0.24)	0.00(−0.01, 0.01)	0.01 (0.44)	0.02 (0.58)	−0.01(−0.06, 0.04)
MAR given Z	−0.08(0.21)	−0.27(0.23)	0.19 (0.18, 0.20)***	−0.67(0.34)	−1.57(0.48)	0.90 (0.86, 0.94)***
n = 450						
MCAR	0.00 (0.12)	0.01 (0.13)	−0.00(−0.01, 0.00)	0.00 (0.18)	0.02 (0.31)	−0.01(−0.03, 0.01)
MAR given X	0.00 (0.12)	0.00 (0.14)	0.00(−0.01, 0.01)	−0.00(0.32)	0.01 (0.31)	−0.01(−0.04, 0.02)
MAR given Z	−0.09(0.12)	−0.29(0.13)	0.20 (0.19, 0.20)***	−0.65(0.18)	−1.50(0.24)	0.86 (0.84, 0.88)***

Biases are reported as mean (SD) for each model, based on 500 simulated replicates. Mean differences are presented with corresponding 95% confidence intervals, and p-values were derived from paired t-tests.
* p < 0.05, ** p < 0.01, *** p < 0.001.
Abbreviations: AIPCCW, augmented inverse probability of complete-case weighting; CCA, complete-case analysis; CI, confidence interval; est, estimate; SD, standard deviation; p, p-value.

Table 3.3: Bias of AIPCCW and IPCCW, and their mean difference, across different sample sizes, proportions of missing data, and missing data mechanisms.

	20% Missing data			85% Missing data		
	AIPCCW Mean (SD)	IPCCW Mean (SD)	Mean difference est (95% CI) p	AIPCCW Mean (SD)	IPCCW Mean (SD)	Mean difference est (95% CI) p
n = 40						
MCAR	0.02(0.40)	0.02(0.43)	0.00(−0.01, 0.02)	−0.04(0.95)	0.01(1.34)	−0.05(−0.15, 0.06)
MAR given X	0.00(0.43)	0.02(0.48)	−0.02 (−0.04, −0.00)*	−0.13(1.20)	−0.06(1.74)	−0.07(−0.20, 0.07)
MAR given Z	−0.14(0.40)	−0.34(0.42)	0.20(0.18, 0.21)***	−0.71(1.07)	-1.78 (1.18)	1.07 (0.96, 1.18)***
n = 150						
MCAR	−0.00(0.20)	0.00(0.21)	−0.00(−0.01, 0.00)	−0.01(0.33)	−0.03(0.50)	0.02(−0.02, 0.05)
MAR given X	−0.02(0.20)	−0.02(0.24)	−0.00(−0.01, 0.01)	0.01(0.44)	−0.01(0.76)	0.01(−0.04, 0.07)
MAR given Z	−0.08(0.21)	−0.27(0.23)	0.19 (0.18, 0.20)***	−0.67(0.34)	−1.59(0.43)	0.92 (0.89, 0.96)***
n = 450						
MCAR	0.00 (0.12)	0.00 (0.13)	−0.00(−0.01, 0.00)	0.00 (0.18)	0.02 (0.28)	−0.02(−0.04, 0.00)
MAR given X	0.00 (0.12)	0.01 (0.14)	−0.01 (−0.02, −0.00)*	−0.00(0.32)	0.02 (0.54)	−0.03(−0.07, 0.01)
MAR given Z	−0.09(0.12)	−0.29(0.13)	0.20 (0.19, 0.20)***	−0.65(0.18)	−1.51(0.22)	0.86 (0.85, 0.88)***

Biases are reported as mean (SD) for each model, based on 500 simulated replicates. Mean differences are presented with corresponding 95% confidence intervals, and p-values were derived from paired t-tests.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
Abbreviations: AIPCCW, augmented inverse probability of complete-case weighting; IPCCW, inverse probability of complete-case weighting; CI, confidence interval; est, estimate; SD, standard deviation; p, p-value.

3.3.2 Missing data in the explanatory variable

Similarly to the previous section, we studied the performance of AIPCCW given various settings and missing data mechanisms mechanisms, but with missing data in the explanatory variable X. In Figure 3.7, it is found that AIPCCW perform with a statistically significance better than CCA and MI, through a bias closer to zero. Additionally, AIPCCW improve and produce a statistically significant difference in bias, when comparing a low level of correlation to a higher level of correlation between the auxiliary variable Z, and the variable with missing values, i.e X. This is further displayed in Figure 3.8, were a decreasing eRMSE can be found as the correlation increases.

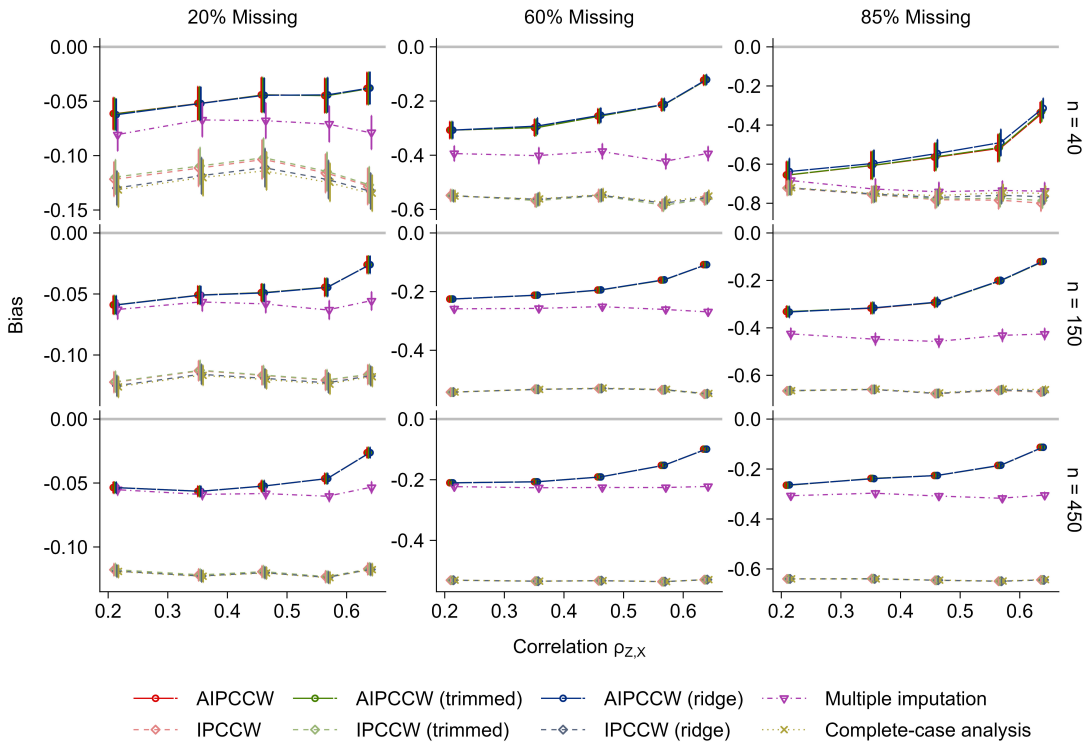

Figure 3.7: Visualisation of bias as a function of the correlation $\rho_{Z,X}$ between the auxiliary and the explanatory variable, comparing different methods of handling missing data at various sample sizes denoted as n, and proportions of data that is MAR given Z in the explanatory variable.

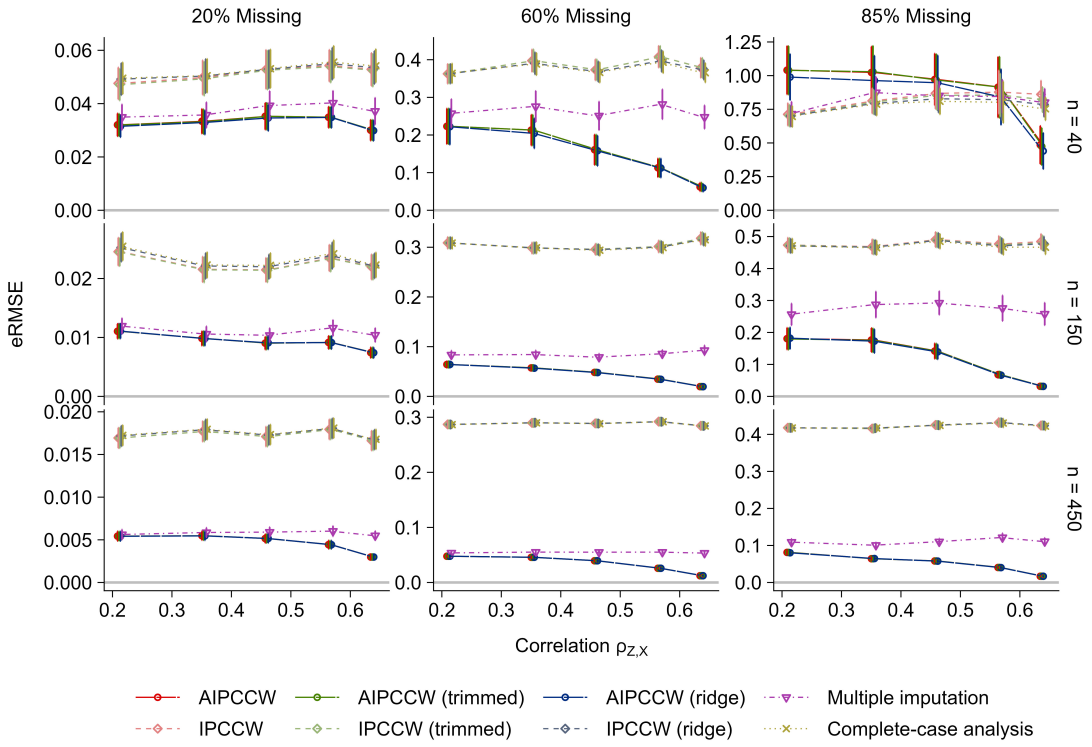

Figure 3.8: Visualisation of eRMSE as a function of the correlation $\rho_{Z,X}$ between the auxiliary and the explanatory variable, comparing different methods of handling missing data at various sample sizes denoted as n, and proportions of data that is MAR given Z in the explanatory variable.

4

Application

This chapter is aimed at analysing and comparing CCA, MI, IPCCW, AIPCCW, based on real-world data.

4.1 Introduction and objectives

In this application study, the following research questions from Chapter 1 are studied and answered: "Can AIPCCW improve estimation and inference in real-world applications in medical research with block-structured missing data?".

4.2 Data

The data were collected to discover the underlying causes and possible treatments to pre-eclampsia, a severe pregnancy disorder which claims 47 000 lives globally each year [8]. Research regarding pre-eclampsia is currently on-going, but encounters challenges with missing data. Results from some of these research studies has been published elsewhere [11, 16], and this thesis is a re-analysis of the same data from the pre-eclampsia obstetric adverse events (PROVE) biobank.

The data were collected in South Africa, where participants included in the study were diagnosed with pre-eclampsia, severe pre-eclampsia, eclampsia, or normotensive controls, i.e. with normal blood pressure during pregnancy. Collected variables included age, weeks and days pregnant (gestation weeks/days) at diagnosis and at tests, height, weight, parity (number of previous births the pregnant participant has experienced), blood-based biomarkers, and other diagnoses such as human immunodeficiency virus (HIV), and polycystic ovary syndrome (PCOS), among other. Additionally, some specialised tests such as brain magnetic resonance imaging (MRI), and tests on cerebrospinal fluid (CSF) to gather information on hormone levels and certain protein levels. However, only a subset of the pregnant participants had specialised test such as CSF analysis done. A descriptive table of some baseline characteristics of the data which the application study in this thesis is build upon can be found in Table 4.1, and a descriptive table of the biomarkers analysed in this thesis is presented in Table 4.2.

Table 4.1: Baseline characteristics of women with preeclampsia or eclampsia and normotensive pregnant controls included in the present analysis derived from the South African PROVE dataset

	Pre-eclampsia/eclampsia n=95	Normotensive n=12
Age	28.8 (6.6)	31.8 (5.9)
Parous	30 (57.9%)	8 (83.3%)
Pre-pregnancy weight, kg	75.8 (18.5)	78.9 (12.9)
Missing	2 (2.1%)	3 (25%)
Height, cm	158.1 (5.7)	157.9 (6.0)
Missing	2 (2.1%)	2 (16.7%)
Birthweight, grams	2182 (826)	3005 (698)
Missing	1 (1.1%)	0 (0.0%)
Days in hospital	7.8 (5.2)	4.0 (1.2)
Missing	5 (5.3%)	1 (8.3%)
Gestation weeks at delivery	35 (3.8)	38 (3.1)
Chronic hypertension	7 (7.4%)	0 (0.0%)
Missing	1 (1.1%)	0 (0%)
Maternal death	0 (0%)	0 (0%)
Intrauterine fetal death	0 (0%)	0 (0%)
Neonatal or infant death	4 (4.2%)	0 (0%)

Binary variables are presented with frequency n (percentage), and continuous variables are presented with mean (SD).
Abbreviations: SD, standard deviation.

Table 4.2: Descriptive statistics of inflammatory biomarkers (IL-1β, IL-6, IL-8, and TNF-α) among women with preeclampsia or eclampsia and normotensive pregnancy controls from the PROVE dataset.

	Pre-eclampsia/eclampsia n=95		Normotensive n=12	
Biomarker	Mean (SD)	Below detection limit n (%)	Mean (SD)	Below detection limit n (%)
IL-1β				
CSF	0.31 (0.26)	49 (52.1%)	0.22 (0.03)	11 (91.7%)
Plasma	3.18 (3.20)	3 (3.16%)	3.85 (4.58)	0 (0%)
IL-6				
CSF	18.0 (53.0)	4 (4.21%)	0.93 (0.39)	0 (0%)
Plasma	24.4 (37.8)	10 (10.5%)	10.6 (10.7)	1 (8.33%)
IL-8				
CSF	267 (563)	0 (0%)	64.2 (20.3)	0 (0%)
Plasma	39.4 (200)	0 (0%)	12.8 (3.11)	0 (0%)
TNF-α				
CSF	3.52 (2.71)	49 (52.1%)	2.84 (1.39)	7 (58.3%)
Plasma	93.6 (51.8)	1 (1.05%)	94.0 (21.8)	0 (0%)

Binary variables are presented with frequency n (percentage), and continuous variables are presented with mean (SD).
Abbreviations: CSF, cerebrospinal fluid; IL, interleukin; SD, standard deviation; TNF, tumour necrosis factor alpha.

4.3 Method

The proportion of missing data ranged from 0.0% to 5.6% across baseline variables and was present in 9.3% of the participants. The data set included five participants with unknown study group, which were excluded from the analyses. For the purposes of this

evaluation, individuals with missing data in biomarker variables were removed from the dataset, and missingness was subsequently introduced according to predefined mechanisms. Contrary to the simulation study in Chapter 3, factors such as sample size and correlation did not vary, as the data and its properties were fixed. The application study were aimed at evaluating AIPCCW, CCA, MI, and IPCCW under varying proportions of missing data, and under differing missing data mechanisms. To accurately evaluate the methods behaviour, several sets of relations between biomarkers were tested. These were chosen based on significant correlation between log-transformed levels of biomarkers in cerebrospinal fluid (CSF) and severity of pre-eclampsia (i.e normotensive, pre-eclampsia, pre-eclampsia with severe features, and eclampsia), as well as significant correlation between log-transformed levels of biomarkers in blood plasma and severity of pre-eclampsia. These biomarkers include interleukin-1β (IL-1β), interleukin-6 (IL-6), interleukin-8 (IL-8), and tumour necrosis factor alpha (TNF-α) [11].

To evaluate the methods, missing data at various proportions was simulated 500 times, after which the methods was applied for each repetition. This was performed for 4 different sets of CSF/blood plasma relations in the data, based on significant correlation between pre-eclampsia and levels of biomarkers in CSF and blood plasma. The 5 different relations and thus models being displayed in Table 4.3.

Table 4.3: Models of biomarkers studied in the application study, where the study group refers to normotensive controls vs. non-normotensive controls.

Response variable	Explanatory variable	Auxiliary variable
IL-1β CSF	Study group	IL-1β plasma
IL-6 CSF	Study group	IL-6 plasma
IL-8 CSF	Study group	IL-8 plasma
TNF-α CSF	Study group	TNF-α plasma

Study group refers to normotensive controls vs. non-normotensive individuals, i.e participants diagnosed with either pre-eclampsia or eclampsia.
Abbreviations: CSF, cerebrospinal fluid; IL, interleukin; TNF, tumour necrosis factor alpha.

4.4 Results

To answer the research question on whether AIPCCW could improve estimation and inference in analyses of real-world data, we studied the performance of AIPCCW in compared to CCA, MI, and IPCCW through evaluation metrics such as bias and eRMSE. In Figure 4.1 we display eRMSE as a function of proportion of missing data for the four (log-transformed) biomarkers IL-1β, IL-6, IL-8, and TNF-α are presented under three missing data mechanisms. Similarly, in Figure 4.2 the performance of the methods are presented through bias. It is found that the performance of in both eRMSE and bias vary greatly given the biomarker studied. It is found that AIPCCW seem to diverge with a large eRMSE, and bias when applied on biomarker IL-6, when 85% data are missing. However, in 4.1 it is found that AIPCCW perform statistically significantly better than MI for biomarkers IL-8 and TNF-α.

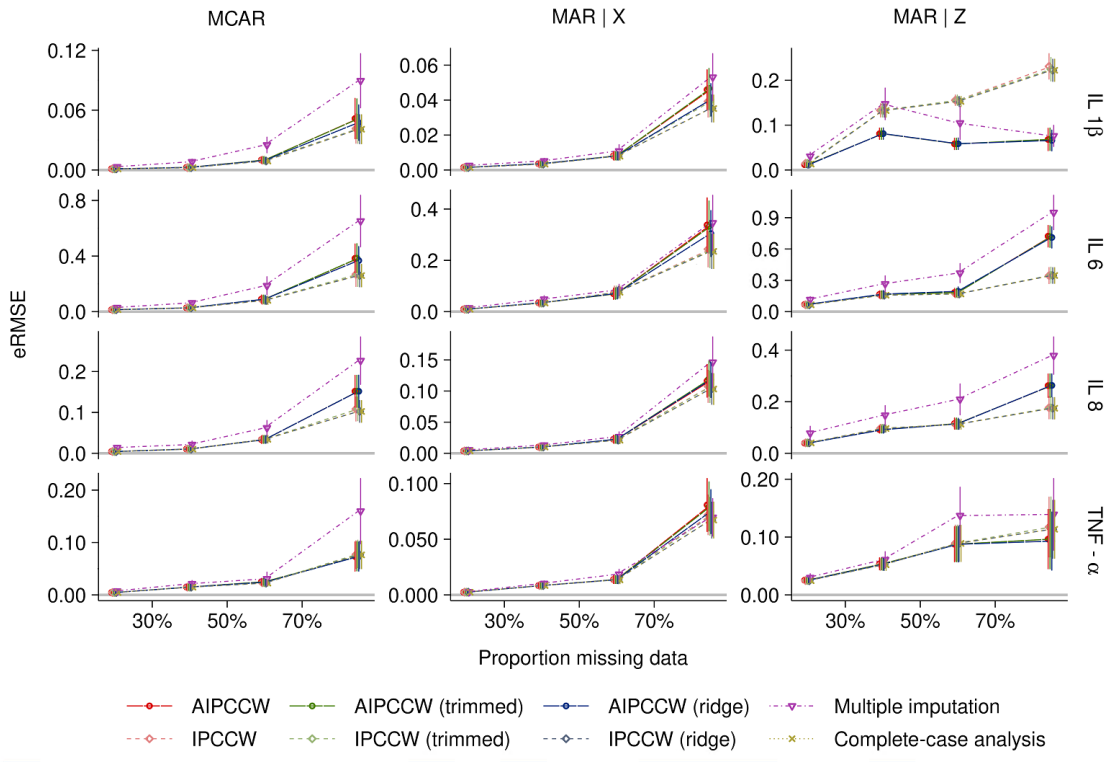

Figure 4.1: Estimation RMSE of methods for (log transformed) biomarkers under different missing data mechanisms as a function of the proportion of missing data.

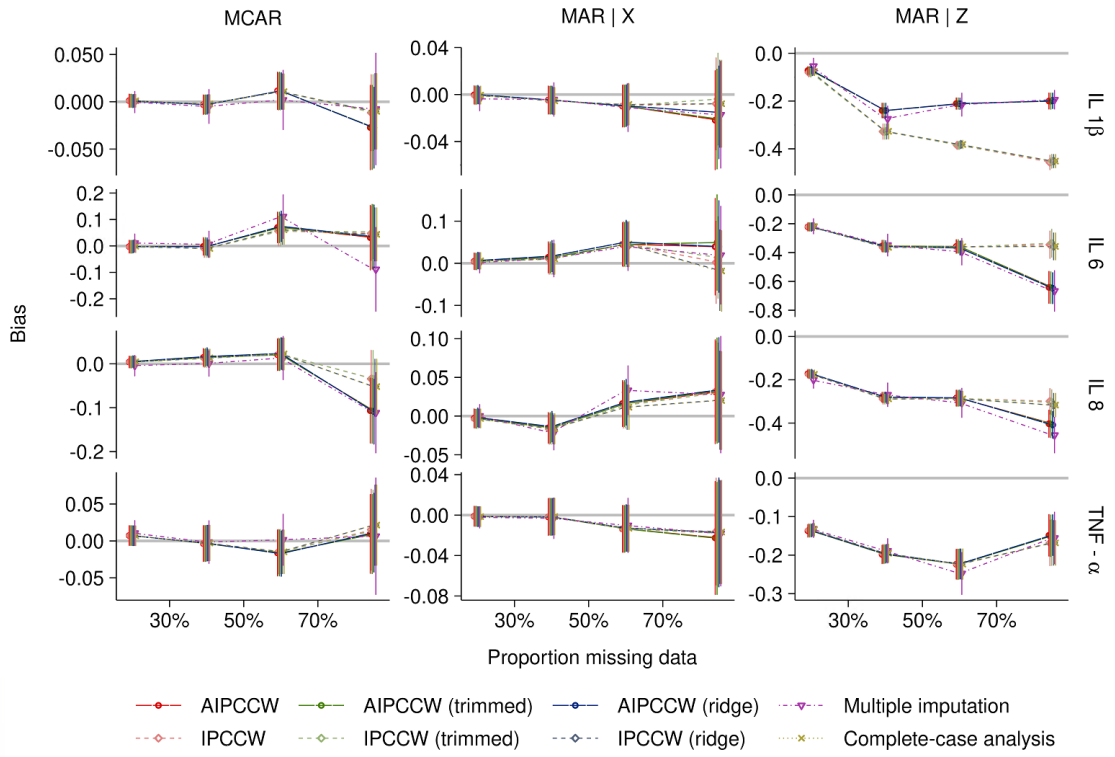

Figure 4.2: Bias of methods for (log transformed) biomarkers under different missing data mechanisms as a function of the proportion of missing data.

5
Discussion

In this chapter a short description of the study is presented and its main findings. This will be followed by Section 5.2 Interpretation and explanations, Section 5.3 Clinical and practical implications, and Section 5.4 Methodological considerations, limitations, and future work. The conclusion and take home message from this thesis are presented in the end of this chapter, in Section 5.5 Conclusion.

5.1 Main findings

This thesis evaluated the performance of augmented inverse probability of complete-case weighting compared to the three benchmark methods complete-case analysis, multiple imputation, and inverse probability of complete-case weighting, in different scenarios. It was found that AIPCCW performed well, and significantly improved given a higher correlation between the auxiliary variable and the variable with missing data, but did only significantly perform better than CCA, MI, and IPCCW in a small minority of scenarios. The four methods were compared through evaluation metrics such as bias and eRMSE, in different scenarios on both simulated data, and on real-world data. In the simulation study, the different scenarios included combinations of factors with multiple levels such as missing data mechanisms, sample sizes, proportion of missing data, and level of correlation between auxiliary variable and variable with missing values. In the study where real-world data was used, the different scenarios included combinations of factors with multiple levels such as missing data mechanisms, and proportion of missing data.

Across the simulation scenarios AIPCCW was only found to perform better than CCA, MI, and IPCCW when data was MAR conditionally on the auxiliary variable Z. Yet AIPCCW produced significantly lower eRMSE than any other method in scenarios with 85% missing data when the sample size were at most 150, and the correlation between the auxiliary variable and the variable with missing data were large, regardless of missing data mechanism studied, as seen in Figure 3.4. On the other hand, in Figure 3.5 we do only find a narrower confidence interval when data are MCAR, meanwhile when data are MAR conditionally on X we find that AIPCCW fail to be unbiased when the sample size is 40 including 85% missing values. Given the simulation results, it is thus only in scenarios where data are MAR conditionally on Z that AIPCCW perform better than CCA, MI, and IPCCW in both eRMSE, and bias.

When evaluating the four methods on real-world data, AIPCCW did not perform significantly better than any other method. However, we found CCA and IPCCW to produced a significantly lower eRMSE than MI when data were MCAR for all biomarkers in Figure 4.1, meanwhile MI produced a significantly lower eRMSE than CCA and

IPCCW for one of these biomarkers when 85% missing data were MAR conditionally on Z. AIPCCW is close to producing a significantly lower eRMSE than MI when real-world data are MCAR, but does not, as can be seen in Figure 4.1. Note that a significant difference in bias are only presented when studying biomarkers that had substantial levels of missing data that were MAR conditionally on Z, which is displayed in Figure 4.2.

5.2 Interpretation and explanations

The main findings from this thesis presented above, are in this section interpreted and explained. We found that AIPCCW applied on real-world data only performed better in both bias and eRMSE than CCA and IPCCW in one out of nine scenarios. A possible explanation would be a too low correlation between the auxiliary variable and the variable with missing values. This can, in real world data, not be adjusted or effected, yet in this case the choice of a binary study group (referring to normotensive controls vs. non-normotensive individuals) may have effected the performance of AIPCCW. This is due to the choice of biomarkers were based on significant correlation between the biomarker and the four groups of pre-eclampsia, severe pre-eclampsia, eclampsia, and normotensive controls. The reason for implementing a binary variable were due to interest in the two special cases described in Section 2.3.4 which this thesis aimed to focus on, and a categorical explanatory variable would introduce a model that fall outside the project's scope. It is also interesting to have found IPCCW to perform well in the real-world data, which could be due to the large separation in participants in the two groups, as inverse probability is a common tool to use in scenarios with unequal sampling fractions which we are presented with. We also found IPCCW to perform at a similar level as CCA in bias and eRMSE, in both real-world data, and a majorty of scenarios on simulated data. This is likely due to IPCCW being a method closely related to CCA, but with the additional weights based on the probability of the participant having complete data and thus the performance of IPCCW and CCA is roughly equal when the weights are roughly equal. This could for example occur more often in scenarios where the missing data are MCAR.

Furthermore, in the case of missing data mechanism MAR conditionally on Z, MI produces a strong bias as can be seen in Figure 3.6. One possible cause is that the given scenario violates the assumption of multiple imputation, as the data are missing conditionally on Z, but Z is not included in the linear regression model of interest $Y|X$ and thus not included in the model when imputing using MI. Thus MI fails as in a case with substantial levels of missing data and a small sample size.

Additionally, we found AIPCCW to improve in both bias and eRMSE, when subjected to a higher level of correlation between the auxiliary variable and the variable with missing data. This is due to the formulation of the approach, as AIPCCW utilisation of the additional information (second term in (2.5)) becomes more effective. In some missing data mechanisms AIPCCW perform at a similar level as other methods, meanwhile in other missing data mechanisms AIPCCW outperform the benchmark methods CCA, MI, and IPCCW. However, it is not possible in real-world situations to prove whether missing data is missing at random, or missing not at random.

No statistically significant difference was found between analyses performed with and without weight stabilisation in AIPCCW. This implies that the approach on weight stabilisation method has limited impact on the outcome. As already noted, there is lim-

ited literature on the practical use of AIPCCW and a possible explanation to why the approach of AIPCCW has not been practically utilised is that other researchers have uncovered some issues with the approach. This possible explanation is supported by the fact that the performance of AIPCCW, and the possible gain of the use of AIPCCW, were dependent on the data, and the missing data mechanism. Additionally, multiple articles were published regarding AIPCCW between 1990 and 2010, but for the past 15 years, only a few new articles has, to my knowlegde, been published regarding AIPCCW (that was not implemented in i setting involving causal inference), and the interest in this approach seem to have declined.

5.3 Clinical and practical implications

AIPCCW is a method that is hard to generalise in a programming language, and it would need to be modelled for each unique set up. AIPCCW would need a complex modelling in cases where several variables has missing data, as auxiliary variables need to be correlated to several variables. Results from the simulation studies also suggests that a high correlation between the auxiliary variable and the variables with missing data should preferably be as high as possible, meanwhile such strong correlations in real-world application cannot be guaranteed or affected.

5.4 Methodological considerations, limitations, and future work

The study is subject to assumptions about auxiliary variables and the missing data mechanism, which may limit generalisability. Computational feasibility imposes restrictions on model complexity, sample sizes, and simulation scenarios, and the analysis focuses on linear models. The simulation study was limited to standard statistical evaluation metrics, excluding comparison of means and adjustments for confounders, which is a possible future project. Furthermore, this study was limited to two special cases of missing data to implement algorithms previously implemented [5], which did not include multiple explanatory variables and were thus the statistical models in this thesis were not adjusted for that. On the other hand, the real-world data had significant correlations only when multiple categories of diagnoses were defined, and the performance of AIPCCW could for this reason be impacted to the worse due to the limitation of only one explanatory variable. Further research on the performance of AIPCCW is needed in scenarios involving nonlinear models, missing data in multiple variables, or the influence of other parameters. In addition, studies on the practical generalisation of AIPCCW would be valuable.

However, more research needs to be carried out, as this study was limited to two special cases of missing data, and a specific real-world data set, and that an other implementation in other modelling scenarios could yield different results.

5.5 Conclusion

The performance of AIPCCW significantly improved with a higher correlation between the auxiliary variable and variables with missing data compared to a lower level of correlation, apart from scenarios with small sample sizes. It was found that eRMSE and bias for AIPCCW decreased in the same rate as CCA, MI, and IPCCW as the sample size increased. AIPCCW was found to outperform CCA, MI, and IPCCW when the missing data mechanism was MAR conditionally on Z. Additionally, AIPCCW generally performed comparable to MI on real-world data, but achieved a lower eRMSE than MI when the data were MAR conditional on Z. However, this thesis implemented two special cases where AIPCCW did not in a majority of scenarios outperform the benchmark methods, and the potential complex modelling of AIPCCW dependent on how many variables included in the model, which variables contain missing values and are strongly correlated to one or several auxiliary variables, could pose a challenge. The potential challenge with a complex model, and the questionable gain in performance, one could in real-world scenarios apply Occam's razor and favour simpler models with similar performance to that of the benchmark methods. AIPCCW is a method with strong theoretical properties, and has shown potential to be a good method in this thesis if a high correlation between the variable with missing values and the auxiliary variable exists. However, this study does not necessarily recommend it over more common methods, but rather recommends that it be implemented alongside other methods in the same study. That said, further research is needed.

Bibliography

[1] A. J. Shattock, H. C. Johnson, S. Y. Sim, A. Carter, P. Lambach, R. C. W. Hutubessy, K. M. Thompson, K. Badizadegan, B. Lambert, M. J. Ferrari, M. Jit, H. Fu, S. P. Silal, R. A. Hounsell, R. G. White, J. F. Mosser, K. A. M. Gaythorpe, C. L. Trotter, A. Lindstrand, K. L. O'Brien, and N. Bar-Zeev, "Contribution of vaccination to improved survival and health: modelling 50 years of the expanded programme on immunization," *The Lancet*, vol. 403, no. 10441, pp. 2307–2316, 2024.

[2] M. Marino, J. Lucas, E. Latour, and J. D. Heintzman, "Missing data in primary care research: importance, implications and approaches," *Family Practice*, vol. 38, no. 2, pp. 200–203, 2021.

[3] P. Mitchell, A. Cribb, and V. Entwistle, "Made to measure: The ethics of routine measurement for healthcare improvement," *Health Care Analysis*, vol. 29, no. 1, pp. 39–58, 2021.

[4] P. Li, E. A. Stuart, and D. B. Allison, "Multiple imputation: A flexible tool for handling missing data," *JAMA*, vol. 314, pp. 1966–1967, 11 2015.

[5] S. Vansteelandt, J. Carpenter, and M. Kenward, "Analysis of incomplete data using inverse probability weighting and doubly robust estimators," *Methodology: European Journal of Research Methods for The Behavioral and Social Sciences*, vol. 6, pp. 37–48, 01 2010.

[6] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," *Statistical Methods in Medical Research*, vol. 22, pp. 278–295, June 2013. Epub 2011 Jan 10.

[7] A. A. Tsiatis, *Semiparametric Theory and Missing Data*. Springer, 2006.

[8] L. C. Chappell, C. A. Cluver, J. Kingdom, and S. Tong, "Pre-eclampsia," *The Lancet*, vol. 398, no. 10297, pp. 341–354, 2021.

[9] M. Fishel Bartal and B. M. Sibai, "Eclampsia in the 21st century," *American Journal of Obstetrics & Gynecology*, vol. 226, no. 2, pp. S1237–S1253, 2022.

[10] L. Duley, "The global impact of pre-eclampsia and eclampsia," *Seminars in Perinatology*, vol. 33, no. 3, pp. 130–137, 2009. Review.

[11] L. Bergman, R. Hastie, H. Zetterberg, K. Blennow, S. Schell, E. Langenegger, A. Moodley, S. Walker, S. Tong, and C. Cluver, "Evidence of neuroinflammation and blood-brain barrier disruption in women with preeclampsia and eclampsia," *Cells*, vol. 10, no. 11, p. 3045, 2021.

[12] R. K. Ross, A. Breskin, and D. Westreich, "When is a complete-case approach to missing data valid? the importance of effect-measure modification," *American Journal of Epidemiology*, vol. 189, pp. 1583–1589, 06 2020.

[13] S. G. Dashti, K. J. Lee, J. A. Simpson, I. R. White, J. B. Carlin, and M. Moreno-Betancur, "Handling missing data when estimating causal effects with targeted

maximum likelihood estimation," *American Journal of Epidemiology*, vol. 193, pp. 1019–1030, 02 2024.

[14] A. E. Valojerdi and L. Janani, "A brief guide to propensity score analysis," *Medical Journal of the Islamic Republic of Iran*, vol. 32, p. 122, 2018. eCollection 2018.

[15] N. Solomon, Y. Lokhnygina, and S. Halabi, "Comparison of regression imputation methods of baseline covariates that predict survival outcomes," *Journal of Clinical and Translational Science*, vol. 5, no. 1, p. e40, 2020.

[16] V. Bucher, O. T. Herrock, S. Schell, J. Visser, H. Imberg, J. Burke, H. Zetterberg, K. Blennow, S. P. Walker, S. Tong, J. Ek, C. Cluver, and L. Bergman, "Blood-brain barrier injury and neuroinflammation in pre-eclampsia and eclampsia," *eBioMedicine*, vol. 116, p. 105742, 2025.

A

Supplemental results

Table A.1: Comparison of bias in the performance of AIPCCW without a weight stabilisation method (baseline) against AIPCCW with trimmed weights, and AIPCCW with weights estimated using Ridge regression.

	AIPCCW† Mean (SD)	Mean difference ‡ est (95% CI) p-value	Mean difference § est (95% CI) p-value
20% Missing data			
n = 40			
MCAR	0.02 (0.41)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	-0.00 (0.44)	-0.00 (-0.00, 0.00)	-0.00 (-0.01, 0.00)
MAR given Z	-0.14 (0.40)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
n = 80			
MCAR	0.00 (0.29)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	-0.01 (0.30)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
MAR given Z	-0.08 (0.28)	-0.00 (-0.00, -0.00)	0.00 (-0.00, 0.00)
n = 150			
MCAR	-0.00 (0.20)	0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
MAR given X	-0.02 (0.21)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
MAR given Z	-0.08 (0.21)	-0.00 (-0.00, -0.00)	0.00 (-0.00, 0.00)
n = 450			
MCAR	0.00 (0.12)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
MAR given X	0.00 (0.13)	0.00 (-0.00, 0.00)	0.00 (0.00, 0.00)
MAR given Z	-0.09 (0.12)	-0.00 (-0.00, 0.00)	0.00 (0.00, 0.00)
n = 600			
MCAR	0.01 (0.10)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	0.00 (0.10)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given Z	-0.09 (0.10)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
60% Missing data			
	AIPCCW† Mean (SD)	Mean difference ‡ est (95% CI) p-value	Mean difference § est (95% CI) p-value
n = 40			
MCAR	-0.02 (0.50)	0.00 (-0.00, 0.00)	-0.00 (-0.01, 0.01)
MAR given X	0.01 (0.60)	0.00 (-0.00, 0.01)	0.00 (-0.01, 0.01)
MAR given Z	-0.48 (0.47)	-0.00 (-0.00, 0.00)	-0.01 (-0.02, -0.00)
n = 80			
MCAR	0.01 (0.34)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
MAR given X	-0.01 (0.43)	0.00 (-0.01, 0.01)	0.00 (-0.00, 0.01)
MAR given Z	-0.49 (0.33)	0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
n = 150			
MCAR	-0.00 (0.24)	0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	0.01 (0.34)	0.00 (-0.01, 0.02)	-0.00 (-0.00, 0.00)
MAR given Z	-0.47 (0.24)	0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
n = 450			
MCAR	-0.00 (0.13)	0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	0.03 (0.23)	0.01 (0.00, 0.02)	0.00 (0.00, 0.00)
MAR given Z	-0.47 (0.14)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
n = 600			
MCAR	0.00 (0.12)	0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	-0.01 (0.21)	0.00 (-0.01, 0.01)	0.00 (-0.00, 0.00)
MAR given Z	-0.48 (0.12)	0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00)
85% Missing data			
	AIPCCW† Mean (SD)	Mean difference ‡ est (95% CI) p-value	Mean difference § est (95% CI)p
n = 40			
MCAR	-0.04 (0.98)	-0.00 (-0.01, 0.00)	-0.00 (-0.02, 0.01)
MAR given X	-0.14 (1.23)	-0.01 (-0.01, 0.00)	-0.01 (-0.02, 0.01)
MAR given Z	-0.75 (1.09)	-0.00 (-0.01, 0.00)	-0.05 (-0.06, -0.03)
n = 80			
MCAR	0.03 (0.54)	0.00 (-0.00, 0.00)	0.00 (-0.01, 0.01)
MAR given X	0.04 (0.64)	0.02 (0.00, 0.03)	0.01 (-0.00, 0.02)
MAR given Z	-0.69 (0.55)	0.01 (0.00, 0.01)	-0.01 (-0.03, -0.00)
n = 150			
MCAR	-0.01 (0.33)	-0.00 (-0.00, 0.00)	-0.00 (-0.01, 0.01)
MAR given X	0.01 (0.48)	0.00 (-0.02, 0.02)	-0.00 (-0.01, 0.01)
MAR given Z	-0.68 (0.35)	0.01 (0.01, 0.01)	-0.00 (-0.01, 0.00)
n = 450			
MCAR	0.00 (0.18)	0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	-0.00 (0.38)	-0.01 (-0.03, 0.02)	-0.00 (-0.01, 0.01)
MAR given Z	-0.65 (0.18)	0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00)
n = 600			
MCAR	-0.00 (0.17)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
MAR given X	0.01 (0.33)	0.01 (-0.01, 0.03)	0.00 (-0.00, 0.01)
MAR given Z	-0.65 (0.16)	0.00 (0.00, 0.00)	-0.00 (-0.00, -0.00)

† AIPCCW with no trimmed weights

‡ Mean difference between AIPCCW with no trimmed weights in comparison to AIPCCW with trimmed weights.

§ Mean difference between AIPCCW with no trimmed weights in comparison to AIPCCW with weights calculated thorguh the use of ridge regression.

Abbreviations: SD, standard deviation; MAR, missing at random; MCAR, missing completely at random; est, estimate.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY