Inverse probability weighted generalised estimating equations for longitudinal data

Andrea Mattsson

Master's thesis
2021:E58

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Inverse probability weighted generalised estimating equations for longitudinal data

LUND
UNIVERSITY

Andrea Mattsson

Master's degree project in Mathematical statistics

Supervisor: Aldana Rosso
Examiner: Anna Lindgren
Centre for Mathematical Sciences, Faculty of Science
Lund University, Sweden
Spring semester 2021

Abstract

Longitudinal study designs, in which variables of interest are observed at multiple time points in a study population, are frequently used in clinical research. Missing data are common in these types of studies. Moreover, in studies investigating a population where the mortality rate is high, data can be truncated by death. To handle missing or truncated data, it is important to investigate the underlying causes to be able to mitigate potentially biased results. By simulating data with different underlying missingness mechanisms, several estimands were investigated in an elderly study population with distinct handling of missing data and truncation by death. Unweighted and weighted generalized estimating equations (IPWGEE) were used to estimate the lung function decline by age. The results suggest that the IPWGEE method is robust when applied to different estimands based on the simulated data set.

Keywords: weighted generalised estimating equations, inverse probability weighting, longitudinal study, missing data, truncation by death, lung function, FEV1

Popular scientific summary

The investigation of lung function in a population plays an important role in diagnosis and assessment of many airway diseases. In addition, an overview of the lung function status can help provide better health-care options and inform medical decisions.

When studying the lung function, the collection of measurements (data) of the lung function from voluntary individuals is a crucial part. Often, several measurements are collected from the same participants over a follow-up period, which allows to explore how the lung function changes over time at both individual and group level. However, sometimes planned measurements cannot be collected due to various reasons. This is referred to as missing data. There are several reasons for missing data, for example if an individual does not show up for a planned follow-up visit, if he or she struggles to perform the measurement or if he or she dies. Missing data have implications for the analysis of the data since information that would help the investigation is lost. Furthermore, the reasons for the data being missing can affect the interpretation of the results.

Gott Åldrande i Skåne (GÅS) is an ongoing study investigating the lung function in elderly participants living in Southern Sweden. The aim of this thesis is to explore the impact of different reasons for missing data on the lung function in elderly individuals, inspired by GÅS.

Among the various approached in how missing data were handled in this thesis, the computations show similar results, indicating that the method is robust under different assumptions. In addition, the results highlight the importance of investigating the reasons why some data were not collected to be able to draw reliable conclusions and make relevant health-care related decisions.

Acknowledgements

First of all, I would like to thank my supervisor Aldana Rosso, Associate Professor in Medical Statistics at Lund University, for her support, guidance and patience through each stage of this project. I am grateful for her trust in me and for inspiring my interest in biostatistics and methodology. I would also like to thank Jakub Hasiec, Data Scientist at the European Medicines Agency (EMA), for his great help with the programming in R. Lastly, I would like to thank colleagues at EMA who kindly have provided constructive comments on the report.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clinical research aims to inform decision-making, improve healthcare options and provide insights regarding human health aspects by establishing a causal relationship between variables of interest. This can be achieved by conducting a study with a suitable design and statistical analyses.

Longitudinal study designs, a design common in clinical research, allow for statistical inference on changes in a variable over time by collecting data from the same study participants across multiple time points [1]. A consequence of collecting measurements from the same participants at several occasions is within-subject correlated data. The generalised estimating equations (GEE) approach is a statistical method for analysing correlated data, such as longitudinal data, by estimating regression parameters in population-averaged models. The within-subject correlated data are handled by incorporating a correlation structure [2, 3].

Longitudinal studies often extend over a long period of time, with numerous follow-up visits, and missing observations are common. Missing data can be defined as unobserved values of variables that were planned to be collected and would be meaningful for the analysis if observed, but for some reason were not collected [4, 5]. If the reason for missing data is associated with the outcome variable of interest, the missingness may lead to biased estimates [1, 6, 7, 8]. There are several methods to handle missing data in the statistical analyses to reduce potential biases. In short, they all have the purpose to either replace the missing value with an adequate guess (imputation methods) or weight the observed data to compensate for the missing data (inverse probability weighting, IPW) [1, 3, 9].

In the design of a study it is important to define one or several estimands. An estimand describes the target of inference by defining what to be estimated and aligning it with the study objective and population. One part in defining the estimand is to specify the handling of events that hinder the collection of variables and therefore may interfere with the interpretation of the study results, also referred to as intercurrent events.

The estimation of an estimand is also impacted by missing data. Missing data can have various causes, for instance when participants drop-out due to a lack of interest to be followed-up, relocation from the area of study inclusion, or failed measurements. These are examples of data that could have been observed, since the participant is alive, but for some reason were not [10]. Another cause of missing data is when participants die during follow-up. This is particularly common when the population of interest is elderly and frail and therefore at a higher mortality risk. However, unobserved data due to death of participants is not a straightforward missing data problem since the variable of interest cannot be observed and hence is undefined. The handling of missing data due to death requires extra attention since imputation or weighting methods would attempt to recover these undefined observations. Alternatively, death can be viewed as an intercurrent event in the estimand framework. Defining death as an intercurrent event and choosing a strategy to handle it will impact interpretation of the results. Depending on the handling of death in the statistical analysis, the inference is based on data from a certain population. For instance, if missing data by death is handled by imputing methods, the following inference would be based on an immortal cohort that is prohibited from dying. The question is if analyses based on such a cohort is of clinical relevance. Therefore, depending on the research question at hand, making a distinction between missing by non-attendance (drop-out) and missing by death when handling missing data in the statistical analyses might provide more relevant results. As an example, it has been proposed that inference about a study population consisting of only alive participants are more suitable when missing data are due to death. This is referred to as mortal cohort (also called partly conditional) modeling [8, 11].

The data analysed in this project are simulated based on an ongoing longitudinal cohort study, Gott Åldrande i Skåne (GÅS), which is part of the Swedish National Study on Aging and Care (SNAC). The overall aim of GÅS is to gain more insight on aspects that affect the general health in an elderly population [12].

One objective of GÅS is to study the lung function change over time. Lung function testing plays an important role for diagnosing obstructive airway diseases such as asthma or chronic obstructive pulmonary disease (COPD), assessing disease severity, and observing treatment responses. The lung function deteriorates naturally with age, and is closely associated with mortality [13, 14]. Lung function tests are performed using a spirometer which quantifies several different lung function outcomes [15]. For the purpose of this investigation, the outcome of interest is *Forced expiratory volume in* 1 *second* (FEV1), which measures the volume of air in litres (L) that a person forcibly can blow out in the first second after taking a full breath [16].

The overall research question addressed in this thesis is how the population-averaged FEV1 change in an elderly study population by age. The main goal of the analysis is to discuss any differences between

the estimands along with possible explanations. We investigate how the estimated population-averaged FEV1 change over time varies for four different estimands by handling the missing data differently based on its cause: missing by drop-out or missing by death. Two of the estimands are based on an immortal cohort. The difference between these estimands rests in the amount of information on the cause for missingness. The two other estimands are based on a mortal cohort in the sense that they only include data from alive participants. One of these mortal cohort estimands is based on unobserved data which can only be known for simulated data. The purpose of this estimand is to serve as a benchmark to which the other three estimands are compared. We hypothesised that, in younger age groups, the differences between the estimands would be small. In contrast, in older age groups, where the population is expected to be more fragile and have a higher mortality rate, the difference would be larger between the estimated estimands, depending on the amount of information on the cause of missingness used.

The FEV1 change over time is estimated by predicting the population-averaged FEV1 for different age values of the study population. The statistical method implemented is inverse probability weighted generalised estimating equations (IPWGEE) where the missing data are accounted for by assigning a weight to each observation which is inversely proportional to the predicted probability for a participant to attend a visit [3, 9].

Chapter 2

Background theory

2.1 Longitudinal study design

Longitudinal studies are designed to investigate characteristics of a study population over time by observing an outcome variable and explanatory variables repeatedly for each study subject, hereafter called participants [1]. Longitudinal study designs are found in both clinical trials and observational studies. A longitudinal cohort study is a specific kind of longitudinal study where the participants are sampled based on a common characteristic, for example by age. The main advantage of longitudinal study designs is the possibility to investigate individual or population level change of an outcome variable over time and being able to relate the change to other variables. Longitudinal studies investigating an elderly population, for instance, often focus on ageing of some characteristic over time [1, 3].

When a participant is observed repeatedly over time, the observations from the same participant cannot be considered as independent measurements. For instance, if a participant's lung function is measured at two separate time points, these measurements will be correlated since they originate from the same participant. This is referred to as within-subject correlation. Thus, for valid inference, it is crucial to make use of statistical methods that can account for correlated data in the analysis, such as generalised estimating equations or mixed models [1, 3].

2.1.1 Notation

Suppose a longitudinal study with N enrolled participants planned for J follow-up visits. At each visit j $= 1, \ldots$ J and for each participant i $= 1, \ldots,$ N, observations of the outcome variable, y_{ij}, and P associated explanatory variables x_{ij} are collected. Let y_i be the J $\times$ 1 vector consisting of all observations y_{ij} for each participant i and $X_i = [1, x_{i1}, \ldots, x_{iJ}]^T$ be the (P+1) $\times$ J matrix of explanatory variable values for each participant i. Let the binary variable M_{ij} be the variable indicating if an observation of y_{ij} is missing (observation is missing if $M_{ij} = 1$). Let the binary variable D_{ij} be the variable indicating if participant i

is dead (participant is dead and observation is missing if $D_{ij} = 1$) at visit j. Let A_{ij} be the indicator for missing by drop-out but not death. The set of $\{M_{ij} = 1\}$ is the union of the set of variables $\{D_{ij} = 1\}$ and $\{A_{ij} = 1\}$.

The data analysed are assumed to be in a long data structure, where each row contains the measured variables for each visit and participant. If a variable is unobserved, i.e. missing, the entry for that variable is noted with not applicable (NA) and the missing indicator variable, and possibly the dead indicator variable if applicable, is equal to one. In table 2.1, an example of a long data structure is depicted, for a study with $N = 3$ participants and $J = 3$ visits.

Patient ID i=1, 2, 3	**Visit Nbr** j=1, 2, 3	**Outcome variable** y_{ij}	**Explanatory variable** x_{ij}	**Missing indicator** M_{ij}	**Dead indicator** D_{ij}
1	1	y_{11}	x_{11}	M_{11}	D_{11}
1	2	y_{12}	x_{12}	M_{12}	D_{12}
1	3	y_{13}	x_{13}	M_{13}	D_{11}
2	1	y_{21}	x_{21}	M_{21}	D_{21}
2	2	y_{22}	x_{22}	M_{22}	D_{22}
2	3	y_{23}	x_{23}	M_{23}	D_{23}
3	1	y_{31}	x_{31}	M_{31}	D_{31}
3	2	y_{32}	x_{32}	M_{32}	D_{23}
3	3	y_{33}	x_{33}	M_{33}	D_{33}

Table 2.1: Example of a long format data set.

The response variable is assumed to drive the missingness, meaning that the covariates are only missing if the response variable is missing.

2.2 Statistical methods for longitudinal data

In medical research, the general linear model (GLM) is a commonly used regression model for estimating associations between an outcome variable and explanatory variables indicating important characteristics of a study subject [2, 17, 18]. Given the setting described in section 2.1.1 with one visit (J=1) visit, the fitted linear predictor for the GLM is given by

$$g(\mu_i) = X_i\hat{\beta}$$

where $\hat{\beta} = (\hat{\beta}_0, \ldots \hat{\beta}_P)^T$ is a (P+1) $\times$ 1 vector of estimated regression coefficients and g is a link function. In short, given an assumed underlying distribution of the outcome variables y_i, the relationship between μ_i and the explanatory variables is modelled using the link function g and the estimated regression coefficients $\hat{\beta}$. The link function allows for inference for data with a range of different underlying error distributions such as normal, Poisson or binomial.

The regression coefficients are estimated using maximum likelihood (ML) methods which assume an underlying probability distribution for the response variable. One of the key assumptions of the GLM is that the measurements are independent observations of a random variable [17, 18].

However, as previously mentioned, longitudinal data have a within-subject correlation which needs to be adequately accounted for in the model. If the repeated measurements for each participant are correlated, the independence assumption of the standard GLM does not hold. Instead, a method that can handle correlated data needs to be applied.

The two most commonly used statistical methods for longitudinal data modeling are the random effects models (also called mixed models) [19] and the generalised estimating equations (GEE) [20].The main difference between these methods is the target of estimation. The random effects model estimates subject-specific changes and is useful when the target of estimation is individual trajectories of the outcome over time. The GEE, on the other hand, fits a marginal model which models population-averages. Marginal models describe how the population average of an outcome variable depends on the explanatory variables, as opposed to the conditional model, e.g. a randoms effects model, which describes how the outcome variable depends on the explanatory variables for each subject [2]. Due to this feature, GEE cannot be used to model individual trajectories over time for an outcome variable, but it is useful when the target of estimation is the average population trajectories over time.

In this thesis, the GEE approach will be considered.

2.2.1 Generalised estimating equations (GEE)

Liang and Zeger (1986) introduced GEE as an extension of GLM to handle within-subject correlated data in marginal models. GEE is a semi-parametric multivariate method to fit a marginal model to within-subject correlated outcomes y_{ij} and explanatory variables vector x_{ij}. It estimates population-averaged outcomes by using a quasi-likelihood method (QL) which is similar to the ML approach, however without any underlying distributional assumptions regarding the outcome variable [20].

As opposed to the ML method, an underlying joint distribution of the outcome variable is not assumed in the QL approach [1]. Instead, the variance of the outcome variable, $v(y_i)$ is assumed to be related to the expectation of y_i, μ_i. This relationship is defined by the assumed marginal distribution of each y_i. The within-subject correlation is accounted for using a working correlation matrix describing the pairwise correlation for each within-subject measurement [20].

Analogous to standard GLM, GEE can model different types of variables such as categorical or continuous data. In contrast to standard GLM, the outcomes and errors are not assumed to be independent,

and homogeneity of the variance is not required [2].

An advantage of GEE is that the working correlation matrix does not need to be perfectly pre-specified to provide sensible coefficient and standard error estimates [21].

Suppose a longitudinal study with the setting described in section 2.1.1. For each participant i =, . . . , N, the linear predictor of y_{ij}, with respect to the marginal distribution for each j, is given by

$$g(\mu_{ij}) = X_i\hat{\beta}$$

where $\hat{\beta}$ is the P $\times$ 1 vector of regression coefficients, with true value β, and g is the link function. For continuous outcome variables, as in this simulation study, the identity link function is applied [20].

Assume that y_{ij} has the probability density function

$$f(y_{ij}) = \exp\left\{[y_{ij}\theta_{ij} - b(\theta_{ij}) + c(y_{ij})]/\phi\right\}$$

where b and c are functions specifying the distribution. θ_{ij}, called the canonical parameter of location, is a function of μ_{ij} and ϕ is the parameter of scale and related to the variances [22]. Moreover,

$$\mu_{ij} = E(y_{ij}) = b'(\theta_{ij})$$

and

$$v_{ij} = Var(y_{ij}) = b''(\theta_{ij})\phi,$$

also known as the first and second moments, respectively.
Define the working correlation matrix $R_i(\alpha)$ for $y_i = (y_{i1}, \ldots, y_{iJ})$ (each subject outcome cluster) with entries

$$\text{corr}(y_{ij}, y_{ik}) = \begin{cases} 1, & j = k \\ \alpha_{jk}, & j \neq k \end{cases}$$

and let $A_i = \text{diag}(v_{ij})$. The variance matrix of y_i is defined by

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

where $\hat{\beta}$ is estimated by solving the general estimating equations defined by

$$U(\beta) = \sum_{i=1}^{N} D_i^T V_i^{-1}(y_i - \mu_i) = 0$$

where $D_i = \frac{\partial \mu_i}{\partial \beta^T}$

The working correlation matrix $R_i(\alpha)$ specifies the variance function and pairwise correlation pattern for each collection of outcome variables per participant. The structure of it is usually unknown but via an inital guess, the correlation stucture is estimated by an iterative procedure, often the Newton Raphson algorithm. When there are relatively few repeated measurements but many study subjects, the unstructured working correlation matrix can be used, for example for measurements per study subject [23]. An unstructured working correlation matrix is defined by unique entries except for the diagonal with entries equal to one such as

$$R_i(\alpha) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} & \alpha_{15} \\ \alpha_{21} & 1 & \alpha_{23} & \alpha_{24} & \alpha_{25} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{34} & \alpha_{35} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 & \alpha_{45} \\ \alpha_{51} & \alpha_{52} & \alpha_{53} & \alpha_{54} & 1 \end{bmatrix}$$

where $\alpha_{jk} = corr(y_{ij}, y_{ik})$ for visit $j \neq k$. When $j = k$, the correlation is equal to 1.

Liang and Zeger (1986) showed that $\hat{\beta}$ is consistent even if the working correlation matrix is misspecified. However, the standard error, will not be correctly estimated. Therefore, a robust estimator of the variance was introduced, with the underlying idea to use the empirical evidence to update the covariance and to adjust the standard errors. This sandwich estimator is defined as

$$N\left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{N} D_i^T V_i^{-1} \text{cov}(y_i) V_i^{-1} D_i\right) \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1}$$

where the parameters are updated over the iterations.

2.3 Missing data

As previously mentioned, there are various possible reasons for why the data were not collected. In clinical research, missing data occur when, for instance, a study participant does not attend a visit and hence the variables of interest were unobserved. Data that is not observed due to death of a participant, for example the measurement of lung function are undefined. Consequently, distinguishing missing data caused by non-attendance and by death in the statistical handling might provide more relevant results, as compared to making no distinction, depending on the research question at hand [4, 8].

Missing data can be grouped into different patterns which indicate if the missingness follow a certain structure. For example, in a longitudinal study setting, missing data can be monotone missing, meaning that if an observation for participant i at visit j, y_{ij}, is missing, all planned subsequent observations y_{ik},

k>j from participant i are missing as well. An example of monotone missing is when a study participant drops out completely from the study, which is called loss-to-follow-up. In contrary, if the participant returns for a subsequent visit k>j after having missed visit j, the missingness follows an intermittent, or arbitrary, missingness pattern [1]. In table 2.2, these two different missingness patterns are illustrated.

Visit	**Missing indicator** M_{ij} Monotone missing	**Missing indicator** M_{ij} Intermittent missing
1	0	0
2	0	0
3	1	1
4	1	0
5	1	0

Table 2.2: Two different missingness patterns, monotone and intermittent missing.

2.3.1 Missing data mechanisms

Missing data can be categorised based on different underlying mechanisms which examine how the missingness of the response variable is related to collected variables or other external factors. Rubin (1976) divided the mechanisms for missing data and the related underlying assumptions into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random MNAR [5].

The MCAR assumption means that the probability for an observation to be missing is independent of observed or unobserved data. Hence, MCAR data comprise of a random sample of the data that were supposed to be collected.

MAR assumes that the missingness is dependent on observed variables. For example, suppose FEV1 measurements are to be collected in elderly participants. Possibly, some of the participants suffer from dementia and cannot understand instructions, and therefore have difficulties to perform a successful spirometry measurement, resulting in missing data. Since one of the variables collected is whether a participant is diagnosed with dementia, the missingness mechanism can be assumed to be MAR. MAR implies that, conditional on the observed variable upon which the missingness is dependent, the missingness is MCAR. This means that, within the subgroup of participants diagnosed with dementia in the above example, missing data are MCAR.

MNAR means that the reason for missingness depends on unobserved data. As an extension of the above example with patients suffering from dementia, now suppose that the variable indicating whether a participant is diagnosed with dementia is not up-to date. It is possible that participants, especially from an elderly population, suffer from cognitive decline but it is not diagnosed yet. Since the missingness

depends on an unobserved variable, this is an example of data that are MNAR [5].

2.3.2 Consequences of missing data

Missing data are an unavoidable issue in most medical research and can be a restraint when interpreting longitudinal data. As already mentioned, ignoring missing data in the statistical analysis may introduce selection biases. Selection bias is a type of bias that may arise in the selection of a study sample. Missing data can be viewed as a type of selection of data and if this selection is not completely at random, the study sample risks not being representative for the target population. If, for instance, only the complete cases are analysed, and therefore the missing data are ignored, selection bias is a risk since completers may or may not be representative of the target population. The potential bias introduced by having missing data can be mitigated by investigating the reason for missingness and addressing the underlying missing data mechanism in the statistical analysis [1, 6, 7, 8]. Figure 2.1 illustrate how missing outcome data Y related to the value of the explanatory variable A bias the results of a regression model.

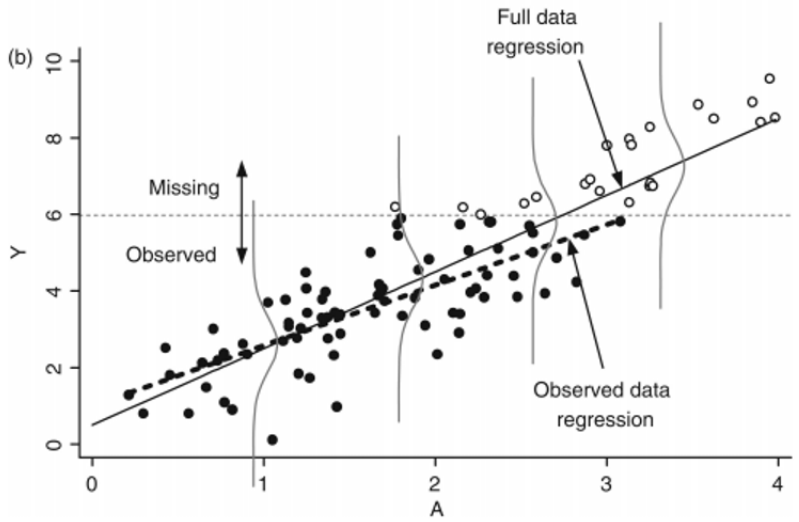


Figure 2.1: Illustration of how missing outcome data Y (y-axis) related to the value of an explanatory variable A (x-axis) bias the results of a regression model [24].

Another limitation for the statistical analysis is the reduced sample size, since data are missing, which implies an increase of variance of the estimates and leads to loss of power in hypothesis testing [25].

2.3.3 Statistical methods to handle missing data

Although it is preferable to avoid missing data in a study by, for instance, encouraging participants to attend the follow-up visits, missing data are common in most studies. To avoid selection bias caused by missing data, the missingness needs to be addressed in the statistical analyses. For this purpose, several methods exist.

A naïve approach is to only use the complete cases data set and ignore missing data. This is called complete cases analysis (CCA). Complete cases analysis provides unbiased results if the missingness is MCAR since the missing data are assumed to be randomly distributed in the sample and independent of any variables predictive of the outcome or the outcome itself. Thus, if MCAR, the complete cases sample is representative for the study population. However, if the missingness mechanism is suspected to be MAR or MNAR, complete cases analysis risk providing biased results [5].

Imputation of missing data is an alternative approach where the missing observations are replaced with estimated values based on observed information. An example of a simple imputation method is the last observation carried forward (LOCF) approach. LOCF substitute the missing observation for participant i at visit j by with the last observation at the previous visit $j-1$ for participant i [26]. An advantage of the LOCF approach is that it is easy to implement. However, LOCF risks to provide biased results in an unconservative direction if the outcome of the participants is expected to get worse over time, such as the lung function in an elderly population. A more advanced type of imputation is multiple imputation (MI). Through this approach, each missing observation is imputed by a set of values derived from an underlying distribution of the outcome variable, allowing for variability and uncertainty in the imputed value [27].

Another approach to account for missing data is to weight the existing observation based on available information. This method is called inverse probability weighting (IPW) [7].

In figure 2.2, a flowchart is illustrating what different methods that are suitable under different assumptions about the mechanism and pattern. The IPW method is found to be suitable under the assumption of monotone MAR data.
Regardless of which missing data method is chosen, it is important to record the reason for missingness. This allows for choosing the most appropriate approach for the situation and to distinguish different reasons from each other. For example, sometimes it can be preferable to distinguish missing data from alive participants from missing data caused by death. The reason for this is that if trying to account for observations that do not exist in the analysis, an immortal population will be created that does not allow for death. This will have implications for the interpretation of the analysis [8].

Of note is that the underlying missingness mechanism is an untestable assumption since the reason for missing data always can depend on unobserved variables. However, the use of subject-matter expertise of the possible reasons for missingness to inform data collection and the analysis can help mitigate the impact of missing data on the results.

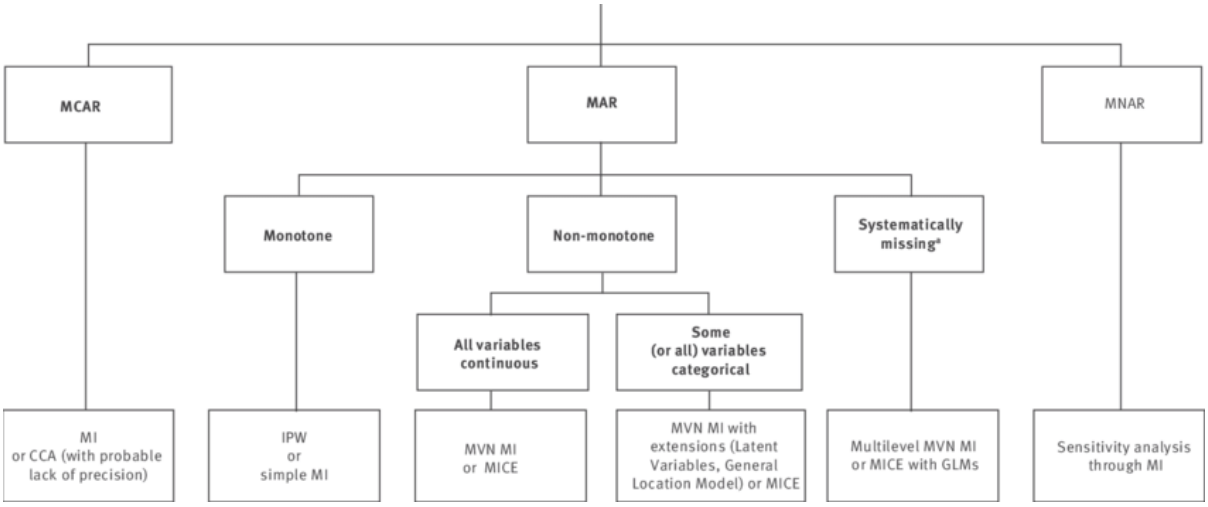


Figure 2.2: Flowchart of appropritate missing data methods under different assumptions. GLM: generalised linear models; MVN: multivariate normal; CCA: complete case analysis; IPW: inverse probability weighting; MI: multiple imputation; MICE: multiple imputation by chained equations; MCAR: missing completely at random; MAR: missing at random; MNAR: missing not at random. [28].

2.3.3.1 Inverse probability weighting

Inverse probability weighting (IPW) is a statistical method that can be used to account for selection bias caused by missing data. It is based on the propensity score approach that was first introduced by Rosenbaum and Rubin (1983) [7, 29].

The propensity score was proposed as a method to handle confounding and selection biases when estimating causal treatment effects in observational studies, which lack the element of randomisation. The propensity score is defined as the conditional probability of receiving a particular treatment, T=t, given the observed data. The propensity score is used to balance treatment groups with respect to observed explanatory variables, meaning that within a group of study subject with the same propensity score, the distribution of the explanatory variables is equal. There are several propensity score-based methods, one of them is IPW. Based on the propensity score, each observation is assigned an IP weight, computed as the reciprocal of the propensity score for a certain treatment t, given that the participant indeed received treatment t [29].

Analogously, IPW can also be applied to reduce selection bias due to missing observations. For this purpose, the variable T indicating which treatment a participant is assigned to, is replaced by the missing variable M, indicating whether the observation is missing or not. In other words, the propensity score is the conditional probability for a study participant of attending a visit, i.e. being observed. The IP weight for each observed data point is inversely proportional to the propensity score of being observed.

The IP weights re-weight the observed data to make the sample representative for the study population. In theory, a pseudo-population is created, with copies of observations to account for missing data, which is twice as large as the population attending the last visit. Because of this, the expected value of each weight is equal to 2. The missing observations are assigned a weight of zero and hence does not contribute to the final estimation model [7].

The purpose of IPW is that the available observations in a sample are weighted to account for the missing observation and make the sample representative for the whole study population. In this way, the information that is lost due to missing observations are compensated for by observations from participants with similar features as the unobserved participants. Each observation is assigned an IP weight based on explanatory variables that are expected to be predictive for the participant to attend the visit.

For each visit, the propensity score for attendance is computed by fitting a logistic regression model with the binary missing variables M_i as outcome and explanatory variables believed to be associated to the missingness [30].

Assume a linear relationship between the logit function the marginal probability of missing the current visit, p_i and the explanatory variables vector S_i for missingness. Then the fitted logistic regression model is given by $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = S_i\hat{\gamma}$ where $\hat{\gamma}$ is the estimated logistic regression coefficients vector.

The predicted probability of being missing is then given by

$$\hat{p}_i = \frac{\exp(S_i\hat{\gamma})}{1+\exp(S_i\hat{\gamma})}$$

The IP weights assigned to attending participants' observations are subsequently given by

$$w_i = \begin{cases} \frac{1}{1-\hat{p}_i}, & M_i = 0 \\ 0, & M_i = 1 \end{cases}$$

Since the missing observations are assigned a weight equal to zero, these participants will only be included in the outcome regression model up to the last visit they attended.

The variance of the IP weights can be large, especially if only a few remaining participants are left for the final visit. Participants with a very low probability of attending will be assigned large weights, yielding in-stable estimates. Stabilised weights were introduced as a method to address the instability. Stabilised weights have the same propensity score in the denominator as the crude weights. However, in addition, the numerator is computed as the marginal probability of attending the current visit, $\hat{q}_i$. These probabilities are equal to the relative frequency of participants current visit, among the participants at-

tending the previous visit, such as

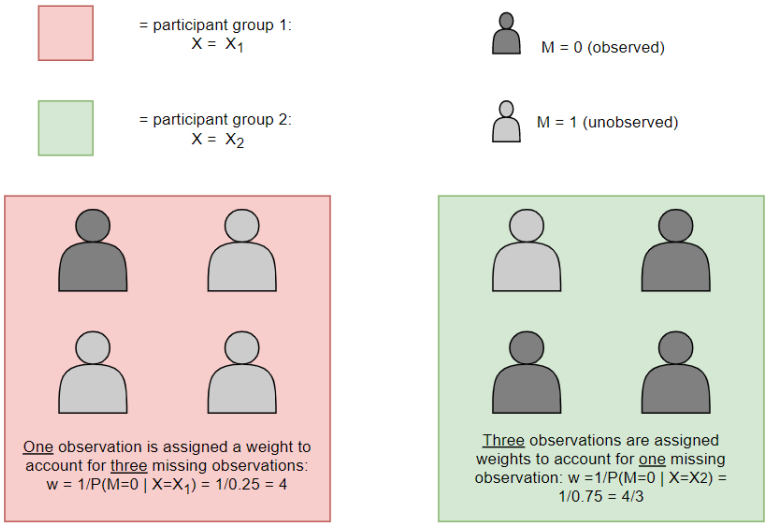
$$q_{ij} = \frac{\#\,\text{participants attending current visit}}{\#\,\text{participants attending previous visit}}$$

thus, for each visit the IP weights are computed as

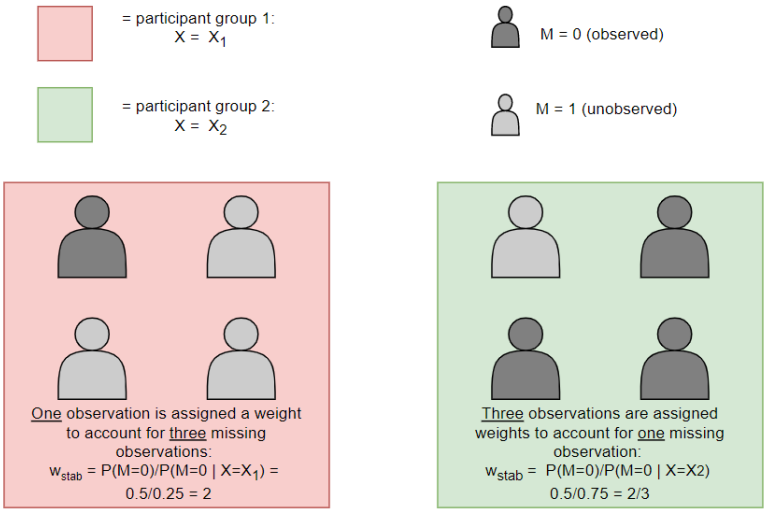
$$w_i^{stab} = \begin{cases} \frac{\hat{q}_i}{1-\hat{p}_i}, & M_i = 0 \\ 0, & M_{ij} = 1 \end{cases}$$

Stabilised weights give a smaller variance. The expected mean of the stabilised IP weights is equal to 1, when the specification of the missingness model is correct [7].

In figure 2.3, a simplified example illustrate how IP weights re-weight the observed data to account for missing observations, for both unstabilised and stabilised weights respectively. The participants are assumed to have one of two possible sets of explanatory variables, X_1 or X_2.



(a)



(b)

Figure 2.3: Illustration of the IPW weighting method for standard IP weights (a) and stabilised IP weights (b). The figure is adjusted from the IPW illustration in [31].

In order for the IPW method to give valid results, the following assumptions need to be met.

Firstly, in the context of IPW for missing data handling, within groups of participants with similar characteristics, defined by the observed explanatory variables, the average outcome is expected to be the equal, regardless which participants generated missing observations. This is referred to as conditional exchangeability. It is an important aspect since IPW re-weight the observations to account for missing observations and for that to increase the representativeness of the observed sample, conditional exchangeability needs to hold in order to give unbiased results [7].

Secondly, there must be a non-zero probability for all participants to provide complete data for all visits. This is referred to as positivity. If this assumption does not hold, it would mean that there is no chance of observing some participants [7].

In addition to these assumptions, the model for the propensity score needs to be correctly specified [30]. This implies that the explanatory variables included in the logistic regression model fitted cannot be mis-specified in order for the IPW to give valid results. It also means that all possible variables that might affect the missingness ideally should be measured. The assumption of no mis-specification of the missingness model is optimally fulfilled by using expert knowledge of probable reasons for missingness and collection of relevant data such that no needed variables are unobserved [7].

An important limitation of IPW in handling missing data is that IPW will not give valid results if the missing-data pattern is intermittent, even if the data are missing at random [7, 32]. Another disadvantage using IPW is that, since it only use the complete cases for each visit, the sample size is still relatively small, since no missing data are imputed [33].

2.4 Weighted generalised estimating equations, IPWGEE

Complete cases analysis using GEE is a valid approach under the MCAR assumpton. However, if the data are missing at random, i.e. the reason for missingness depends on observed explanatory variables (MAR), GEE to not provide consistent results [34]. A combination of IPW and GEE, IPWGEE, was suggested by Robins et al. (1995) as an extension of GEE, to accommodate the issue of data missing at random for correlated data [9].

Let $\Delta_i = \text{diag}(w_{1j}, \ldots, w_{ij}, \ldots, w_{iJ})$ be a J $\times$ J matrix with diagonal elements w_{ij} being the IP weights defined in section 2.3.3.1.

Then the weighted GEE, IPWGEE, is given by

$$U(\beta) = \sum_{i=1}^{N} D_i^T V_i^{-1} \Delta_i (y_i - \mu_i) = 0$$

By multiplying the residual factor $(y_i - \mu_i)$ by the IP weights matrix Δ_i, all observations are weighted in the estimating equations.

IPWGEE is still a complete cases analysis, in the sense that only complete data are analysed and no data are imputed. However, potential selection bias is removed since the pseudo-population created by stabilised IP weighting implies that the selection is distributed randomly in the population [7]. Although IPWGEE may provide valid estimates under the MAR mechanism, an accurate modelling of the weight regression model is crucial for valid inferences [35].

2.5 Estimand framework

A framework for the estimand was introduced in an addendum to the The International Committee for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 guideline on statistical principles for clinical trials. Even though the estimand is focussed on the application on clinical trials, the framework is relevant for observational studies as well. It is also important to note that the estimand framework is mostly discussed in the context of interventional studies in the form of either randomised clinical trials or observational studies [4] [36]. However, the study of interest in this thesis, the GÅS study, is a non-interventional cohort study, where the general health of a study population is investigated over time. Since it is non-interventional, the estimand does not describe a specific treatment effect to be evaluated as in, for example, studying the effect of an experimental treatment compared to placebo on an outcome. Instead, the "treatment" considered is time, measured by age, and its impact on the lung capacity.

The estimand framework is a structured approach to facilitate alignment of study objectives, design, conduct, analysis and interpretation of results. Another aim of the framework is to facilitate the communication of results [4].

An estimand should be specified prior to study start and consists of several attributes. Following the definition of the E9 (R1) Addendum, the features comprise of the target study population, the outcome variable of interest, the treatment to be evaluated, the handling of potential intercurrent events and a population-level summary, see figure 2.4.

Intercurrent events are events that occur after the start of a treatment and that are believed to affect

the collection or interpretation of data. Since these events can affect the research question of interest and the definition of treatment effect, it is important to address how potential intercurrent events will be handled in advance. Hence, this attribute of the estimand is a crucial part. An example of an intercurrent event that can prevent the collection of data is death, given that death itself is not an outcome event of interest. There exist several intercurrent event strategies that are appropriate depending on the type of event. Depending on the strategy, different research questions are addressed, hence the importance to align the strategy to the study objective.

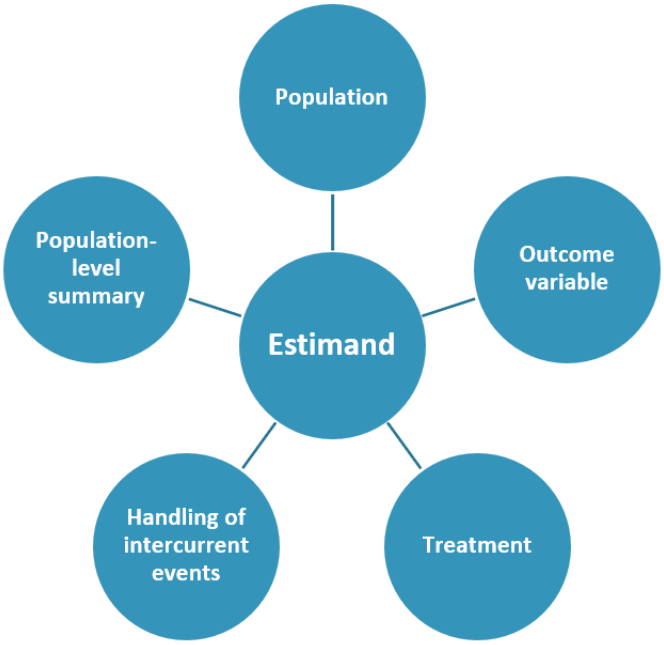


Figure 2.4: The five features that define an estimand.

For the purpose of this thesis, the estimand framework will be utilised to compare different strategies to handle missing data with respect to the handling of death as an intercurrent event.

Chapter 3

Method and data

3.1 Estimands

In order to evaluate how different strategies account for and distinguish reasons for missing data, four estimands will be investigated. The general research question that each estimand aims to target is to study how the lung function in an, in general healthy, elderly population is affected by ageing.

All estimands make use of the IPWGEE estimation model. The estimands differ with respect to distinction of reason for missingness (by non-attendance or death) which in turn has an impact on the target population. In addition, all estimands are investigated using different simulated data sets with respect to the missingness mechanism (MCAR, MAR or MNAR). A description of the estimands is found in table 3.1.

A distinction is made between analyses based on an immortal and a mortal cohort. In the immortal cohort estimand, both missing by non-attendance and by death are handled and adjusted for. This means that undefined observations are implicitly included in the analysis via the method to handle missing data. When using the IPW method to handle missing data, the alive participants are assigned weights to account for dead participants. In the mortal cohort estimand, the analysis is based on data from alive participants, meaning that missing by death is not accounted for in the missing data handling. In this study, four different estimands are used; estimand 1 and 2 use immortal cohorts and estimand 3 and 4 use mortal cohorts.

Estimand 1 makes no distinction between missing by non-attendance or death. The IP weights are computed based on the propensity score of being missing by any cause. Estimand 2, on the other hand, accounts for missing by death differently as compared to missing by non-attendance by computing the IP weights based on the propensity core of being missing by any cause multiplied by the propensity score of being missing by death. This implies that the IP weights assigned to observations to account

for missing by death are generated using additional information on the cause of missingness. Estimand 3 only makes use of data from participants up until they potentially die. This means that no adjustment is made for missing by death data. Estimand 4 is an unweighted complete case estimand based on the simulated data, where the participants that die during follow-up are excluded but without any missing data by non-attendance. Of note is that estimand 4 is unrealistic to be investigated unless the data are simulated. The reason for this is that it builds upon missing data that only are known based on the simulation. The unweighted estimand 4 serves only as a benchmark to which the weighted estimands 1-3 are compared.

Even if the simulated data contain an intermittent missing data pattern, for IPW to give valid inferences, the missingness patterns will be forced to be monotone. This means that if a participant is unobserved, due to any cause, the following observations will be assumed to be unobserved as well.

For each visit $j = 2,\ldots,5$, logistic regression models were fitted with a binary missing indicator vector $M_j = (M_{1j},\ldots,M_{Nj})$, $D_j = (D_{1j},\ldots,D_{Nj})$, or $A_j = (A_{1j},\ldots,A_{Nj})$, as outcome variable and observed data from visit j-1 as explanatory variables such as

$$M_j \sim FEV1_{j-1} + age_{j-1} + sex_{j-1} \tag{3.1}$$

$$D_j \sim FEV1_{j-1} + age_{j-1} \tag{3.2}$$

$$A_j \sim FEV1_{j-1} + age_{j-1} + sex_{j-1} \tag{3.3}$$

where formula 3.1 is used to estimate the IPW weights for estimand 1, both formulas 3.1 and 3.2 are used to estimate the IPW weights for estimand 2, and formula 3.3 is used to estimate the IPW weights for estimand 3.

To estimate the population averaged FEV1 an IP weighted GEE model with FEV1 as outcome variable and age, sex and height as explanatory variables, using the formula

$$FEV1 \sim age + sex + height \tag{3.4}$$

All estimands are defined by the same, treatment, outcome variable and population level-summary such as

- **Outcome variable:** FEV1 to measure lung function
- **Treatment:** Age
- **Population-level summary:** Population-averaged FEV1.

Estimand	Description of features
1	Weighted estimand. **Targeted population:** Immortal cohort **Intercurrent event and its handling:** Death. Dead participants are considered to be missing observations and thereby accounted for by IP weighting. No distinction between missing by non-attendance or missing by death is made.
2	Weighted estimand. **Targeted population:** Immortal cohort **Intercurrent event and its handling:** Death. Dead participants are considered to be missing observations and thereby accounted for by IP weighting. Additional information is provided whether participants is missing by death by fitting two logistic models for IP weights, one for missing and one for death.
3	Weighted estimand. **Targeted population:** Mortal cohort, i.e., inference conditional on being alive **Intercurrent event and its handling:** Death. Participants that die during follow-up only contribute to IPW model up until they die.
4	Unweighted estimand. **Targeted population:** Mortal cohort, i.e., inference conditional on being alive **Intercurrent event and its handling:** Death. Participants that die during follow-up only contribute to IPW model up until they die.

Table 3.1: A description of the four estimands that will be estimated and evaluated.

In the results section, the results for each estimand and cohort size ($N = 50$ and $N = 500$) will be presented. The results presented and analysed consist of the IP weights, the age coefficients for the IPW models together with 95% confidence intervals and predicted population-averaged FEV1 values for different age groups and sex. All estimands will be compared pairwise. However, the comparison to estimand 4 is of primary interest. In addition, the impact of the size of the cohort will be examined by comparing the results for each cohort.

The weights will be examined by investigating the mean of each set of weights for each estimand and visit. Since in theory, the mean of each set of the stabilised weights should be equal to 1, the mean will be explored. In addition, the largest weights for each estimand and visit are presented. Large weights can suggest that only a few elderly frail participants' observations are assigned large weights to account for several dead participants with similar characteristics. This has implication in the interpretation and raise the question of the desired representation of the population.

The age coefficients for the IPWGEE model for each estimands will be studied by point estimate and width of the 95% confidence interval. The reason for highlighting the age coefficient in the results is that age is of primary interest for the research question.

Additionally, in the interest of comparing the full IPWGEE models' performances, the results presented

also entail the predicted population-averaged FEV1 based on the IPWGEE models. Since men and women's FEV1 are rather distinct, the results are stratified by sex. Furthermore, in order to assess the impact of age on the FEV1, the predictions are depicted for the ages 60, 70, 80, 90 and 100 years respectively.

3.2 Simulation of data

The data analysed are simulated to imitate data from the GÅS study with outcome variable FEV1 and explanatory variables sex, age and height, which are all predictors of lung function. The data emulate data from a cohort study with five visits per participants and explanatory variables *age*, *sex* and *height*. Two data sets were simulated with different number of participants, N= 50, henceforth referred to as the *small cohort*, and N = 500, henceforth referred to as the *large cohort*.

The FEV1 data for the first visit are simulated based on the following equations, equation 3.5 for female participants and equation 3.6 for male participants [37].

$$\text{FEV1} = 1.597 + 0.5552 \cdot \text{height}^3 - 0.01574 \cdot \text{height} \cdot \text{age} \tag{3.5}$$

$$\text{FEV1} = 2.081 + 0.5846 \cdot \text{height}^3 - 0.01599 \cdot \text{height} \cdot \text{age} \tag{3.6}$$

where the unit of height is cm and age in years. The only difference between the two formulas are the values of the constants.

FEV1 is estimated to decrease with 22.4 ml annually on average [13]. Hence the FEV1 variables for visit j, $\text{j} = 2, ...5$, are given by

$$\text{FEV1}_j = \text{FEV1}_{j-1}\text{–}0.0225 \cdot (\text{age}_{j-1} - \text{age}_j)$$

For each observation, normally distributed noise, ε is added in order to incorporate extra random variability into the simulations. The sex indicator variable has a binomial distribution, $\text{Bin}(n, p)$, with n equal to the size of the cohort, N, and p equal to $\frac{1}{2}$. The proportions of age of the participants are simulated to reflect the distribution of age in the GÅS study, where the age groups $60-70$ and $80-90$ are the largest. The height variables are normally distributed, $N(\mu, \sigma^2)$.

Three different missing mechanisms are simulated, MCAR, MAR and MNAR via indicator variables $missing_{MCAR}$, $missing_{MAR}$ and $missing_{MNAR}$. In addition a dead indicator variable is simulated using a MAR mechanism, $dead_{MAR}$. The MAR mechanism is simulated to depend on the observed variable *age*. The MNAR mechanism is simulated to depend on the unobserved variable *FEV1* at the current visit.

Moreover, an indicator variable for death is simulated with the underlying assumption of MAR, depending on the observed FEV1 measurements at visit 1. All missing indicator variables missing observations both by non-attendance and by death, meaning that for example $dead_{MARij} = 1 \;\; missing_{MARij} = 1$ for participant i at visit j.

All participants have at least one observed FEV1 measurement from the first visit. This is a requirement for the IPW method to work since the computation of weights is based on the attendance of the last visit.

To reflect the features of real-world data (RWD) where intermittent missing patterns appear, the simulated data also allow for intermittent missing. However, since the different missing mechanisms are simulated using different underlying models, the proportions of intermittent and monotone missing will differ for each missing mechanism. This gives a discrepancy in the number of complete cases for GEE. However, the differences are small, approximately $0.5 - 1\%$ of all planned observations. The simulated variables are described in table 3.2 along with an account for the simulation distributions and dependencies.

Variable	Description of simulation
pat.id	Participant ID number, 1-N
visit	Visit number, 1-5
FEV1	Forced expiratory volume in 1 sec measurements (L). Female participants' FEV1 visit 1: FEV1 = 1.597 + 0.5552 $\cdot$height^3 $-$ 0.01574 $\cdot$ height $\cdot$ age $+ \varepsilon$ Male participants' FEV1 visit 1: FEV1 = 2.081 + 0.5846 $\cdot$height^3 $-$ 0.01599 $\cdot$ height $\cdot$ age $+ \varepsilon$ FEV1 measurements at visit $2-5$: $\text{FEV1}_j = \text{FEV1}_{j-1}$–0.0225 $\cdot$ $(\text{age}_{j-1} - \text{age}_j) + \varepsilon$ where ε is normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 0.1$.
sex	Simulated as a binomial, Bin(n, p), random variable with $n = N$ and probability $p = 0.5$.
age	40% of the participants were uniformly simulated in the age range of $60-70$. 20% of the participants were uniformly simulated in the age range of $70-80$. 30% of the participants were uniformly simulated in the age range of $80-90$. 10% of the participants were uniformly simulated in the age range of $90-102$.
height	Height of participants at first visit, in cm. The height of each participant is assumed to be constant over all visits. Female participants' heights are normally distributed with mean $\mu = 166$ and variance $\sigma^2 = 5$. Male participants' heights are normally distributed with mean $\mu = 180$ and variance $\sigma^2 = 5$. [38]
dead$_{\text{MAR}}$	Indicator variable for missing by death FEV1 observations at random, related to the observed FEV1 measurements at visit 1 such as - participants with a FEV1 observation greater than or equal to the average FEV1 for all participants at the first visit are given a weight uniformly distributed between $0-0.5$; - participants with a FEV1 observation less than the average FEV1 for all participants at the first visit are given a weight uniformly distributed between $0.5-1$.
missing$_{\text{MCAR}}$	Indicator variable for missing FEV1 observations completely at random and missing by death at random.
missing$_{\text{MAR}}$	Indicator variable for missing FEV1 observations at random (MAR) and missing by death at random. The MAR mechanism is related to age at time for missing observation visit such as - participants aged under 80 are given a weight uniformly distributed between $0-0.5$; - participants aged 80 or more are given a weight uniformly distributed between $0.5-1$.
missing$_{\text{MNAR}}$	Indicator variable for missing FEV1 observations not at random (MNAR) and missing by death at random. The MNAR mechanism is related to unobserved FEV1 at time for missing observation visit such as - participants with FEV1 measurement under mean(FEV1) at the current visit (unobserved) are given a weight uniformly distributed between $0-0.5$; - participants with FEV1 measurement equal to or over mean(FEV1) at the current visit (unobserved) are given a weight uniformly distributed between $0.5-1$.

Table 3.2: A description of the simulated data.

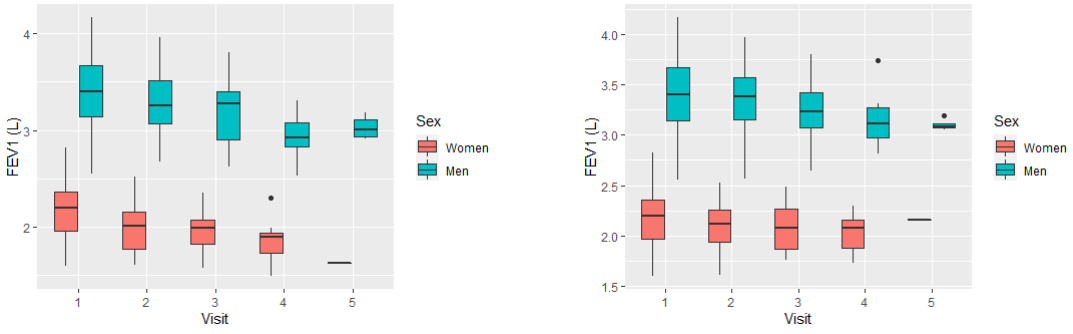
3.3 Data characteristics

In table 3.3, descriptive summary statistics of the simulated study population for the first visit are presented, by cohort. In the small cohort, the average age is 75.80 years as compared to 76.10 years in the large cohort, with both having almost identical standard deviations (sd) of 10.99 and 10.98 respectively. The proportions of female and male participants are equal (50%) in the small cohort as opposed to the large cohort where the proportion of male participants are slightly bigger (53%). The average height for female participants in the small cohort is 167.5 cm in contrast to 166.1 cm in the large cohort. The average height of male participants is almost identical in the small and large cohort, 179.8 and 179.9 cm respectively.

Characteristics	**Small cohort, N=50**	**Large cohort, N=500**
Age in years		
min	60	60
max	99	102
mean (sd)	75.8 (10.99)	76.10 (10.98)
Sex		
Women N (%)	25 (50 %)	235 (47 %)
Men N (%)	25 (50 %)	265 (53 %)
Height in cm		
Women, mean(sd)	167.8 (4.03)	166.1 (5.05)
Men, mean (sd)	179.8 (4.94)	179.9 (5.05)

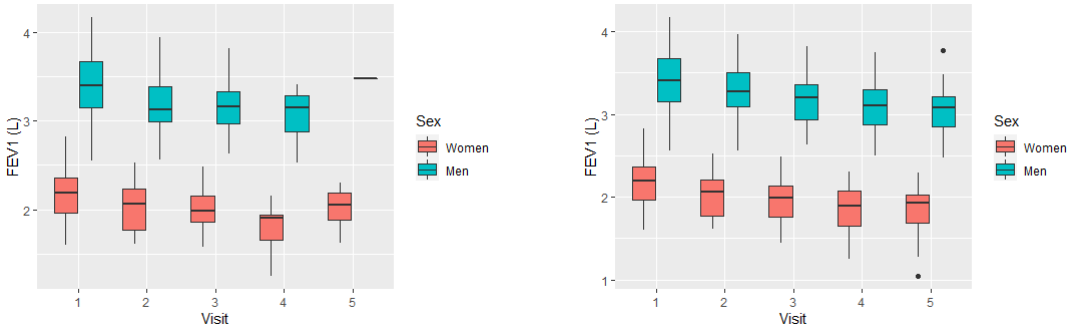
Table 3.3: Data characteristics of the participants at first visit.

In figure 3.1 the observed FEV1 measurements are depicted in boxplots illustrating the distribution of the outcome data by visit 1-5, for each cohort size and missing mechanism.



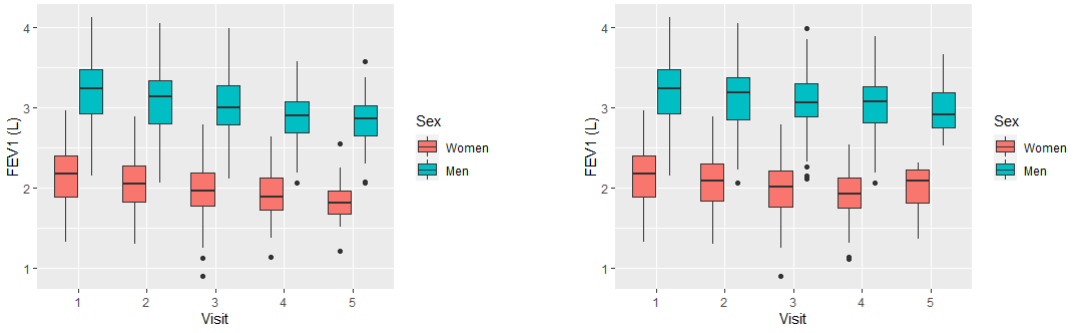
(a) MCAR FEV1 data, N=50

(b) MAR FEV1 data, N=50



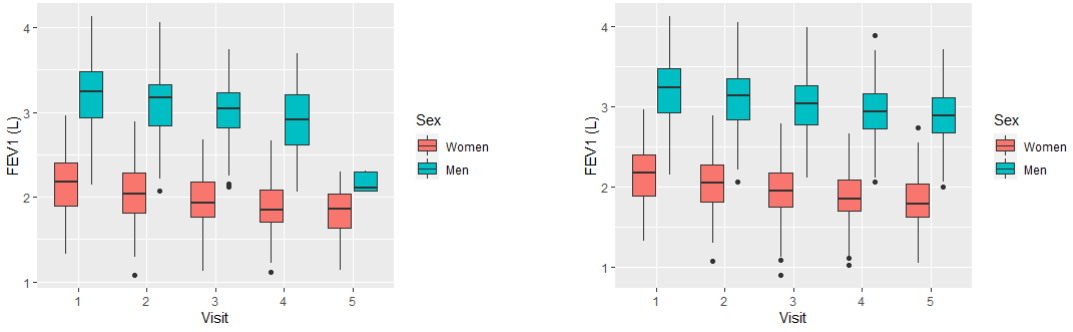
(c) MNAR FEV1 data, N=50

(d) Only missing by death FEV1 data, N=50



(e) MCAR FEV1 data, N=500

(f) MAR FEV1 data, N=500



(g) MNAR FEV1 data, N=500

(h) Only missing by death at random FEV1 data, N=50

Figure 3.1: Boxplots of the 4 different simulated data sets with different missing mechanisms, reasons for misisngness and cohort size.

In table 3.4, the proportions of missing data (both by non-attendance and by death) by visit are tabulated. The missing proportions are equal for all missing mechanisms. The proportion of missing by death out of the missing overall missing proportions is 50% for all visits, except for visit 2 where it is slightly less (43%).

Visit	**1**	**2**	**3**	**4**	**5**
Dead	0%	15%	25%	35 %	45%
Missing	0%	35%	50%	70%	90%

Table 3.4: Proportions of missing data per visit. Both missing by any cause proportions (by non-attendance or death) and only by death proportions are presented.

3.4 Data analysis validation

The IPWGEE estimation has been verified by parallel programming in two separate programming software (R and Stata), warranting for validated results. The different software yield marginally different estimates which can be explained by different underlying algorithms, for example the computation of standard errors. In addition, the package used for the GEE model in R, *geepack*, lacks a method for prediction, whereas Stata offers this possibility with margins command. Therefore, the results presented are evaluated using the rglm and rmargins commands in Stata where the robust standard errors are estimated [39, 40].

Chapter 4

Results

4.1 Small cohort

In table 4.1, the average weights for all estimands are presented, for each visit. Overall, the means of the weights are around 1, suggesting a strength of the IPW method for missing data handling since this is what is expected in theory. Farthest from the expected mean of 1 are the mean of the weights for MAR estimand 1 and 2 for visit 4 with means around 1.30 and the MAR estimand 1 and MNAR estimand 1 and 2 for visit 5, with means around 0.75. In terms of large weights, presented in table 4.2, the estimands that stick out are MAR estimand 1 and 2 for visit 4 with values of around 6. Interestingly, the largest weights for MAR estimand 2 and MNAR estimand 1 are 0.78 and 0.73 respectively.

	Visit number (missing/dead proportions)			
Average weight per visit N=50	**Visit 2 (35%/15%)**	**Visit 3 (500%/30%)**	**Visit 4 (70%/35 %)**	**Visit 5 (90%/45%)**
MAR - Estimand 1	1,01	1,00	1,31	0,84
MAR - Estimand 2	1,01	1,00	1,34	0,76
MAR - Estimand 3	1,00	1,00	0,94	0,90
MAR - Estimand 1	1,03	1,00	0,94	0,98
MAR - Estimand 2	1,03	1,00	0,95	0,97
MAR - Estimand 3	1,02	1,05	0,96	1,10
MAR - Estimand 1	1,00	0,98	1,01	0,71
MAR - Estimand 2	1,00	0,98	1,01	0,75
MAR - Estimand 3	1,02	0,99	1,02	0,86

Table 4.1: Average IP weights for each estimand, by visit, N=50.

	Visit number (missing/dead proportions)			
Average weight per visit N=50	**Visit 2 (35%/15%)**	**Visit 3 (500%/30%)**	**Visit 4 (70%/35 %)**	**Visit 5 (90%/45%)**
MAR - Estimand 1	2,29	1,59	6,47	0,87
MAR - Estimand 2	2,26	1,60	6,89	0,78
MAR - Estimand 3	1,78	1,27	1,99	0,93
MAR - Estimand 1	2,66	1,92	1,42	1,14
MAR - Estimand 2	2,58	1,80	1,35	1,08
MAR - Estimand 3	2,65	2,85	1,49	1,19
MAR - Estimand 1	1,58	1,52	2,74	0,73
MAR - Estimand 2	1,45	1,43	2,49	1,23
MAR - Estimand 3	2,34	1,42	2,05	0,92

Table 4.2: Largest IP weights for all estimands, by visit, N=50.

In figure 4.1, the estimated age coefficients for all estimands are depicted, along with 95% confidence intervals. All estimated age coefficients for the three MCAR estimands are smaller than the estimate for estimand 4, indicating a larger effect of age on FEV1. On the other hand, apart from the MNAR estimand 3, the estimated age coefficients for all other estimands are larger than the estimate for estimand 4. The age coefficient for the MNAR estimand 3 is exactly equal to the age coefficient of estimand. The width of the 95% confidence intervals for the age coefficients are roughly the same, approximately 0.006 units. An exception is the MAR estimand 3 which stands out with respect to the width.

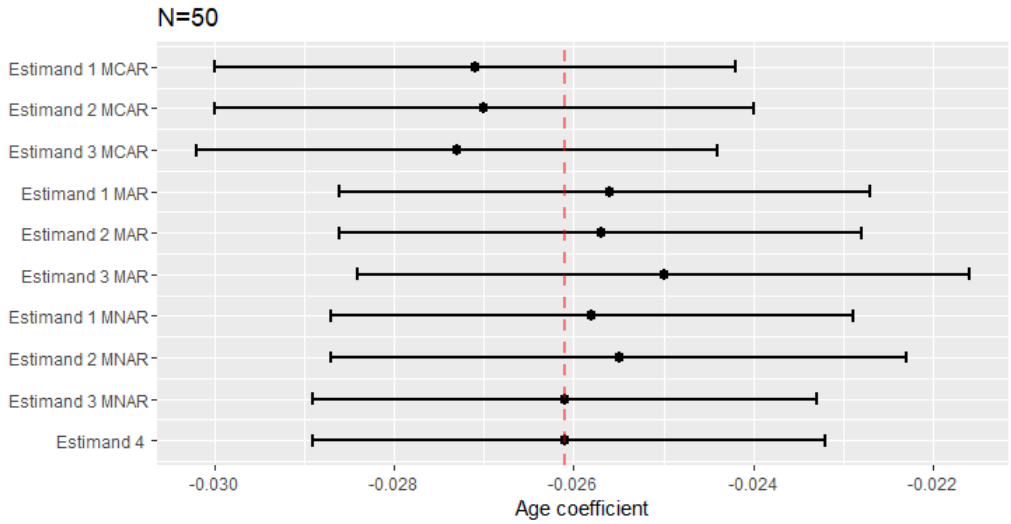


Figure 4.1: Estimated age coefficients from IPWGEE model, for all estimands, N=50.

The predicted population-averaged FEV1 for the small cohort are illustrated in figures 4.2, 4.3 and 4.4. The estimation of all estimands successfully predicts a decline in lung function by age.

The predicted population-averaged FEV1 for 60-year old men in estimand 4 is slightly greater than for the other estimands. This is also in line with the results of the age coefficients, where estimand 4 provided slightly larger estimates in comparison to the other estimands. However, this is not the case for women, suggesting that there might be an imbalance in explanatory variables between sex. Since the underlying formulas for the data are known, it can be concluded that a random imbalance most likely is the reason for these marginally distinct predicted FEV1 values for each sex.

For all estimands, the 95% confidence intervals are narrower for the ages 70, 80 and 90, which can be explained by the larger group of participants in these age groups. The widest 95% confidence interval is shown for 100-year olds, for both sexes.

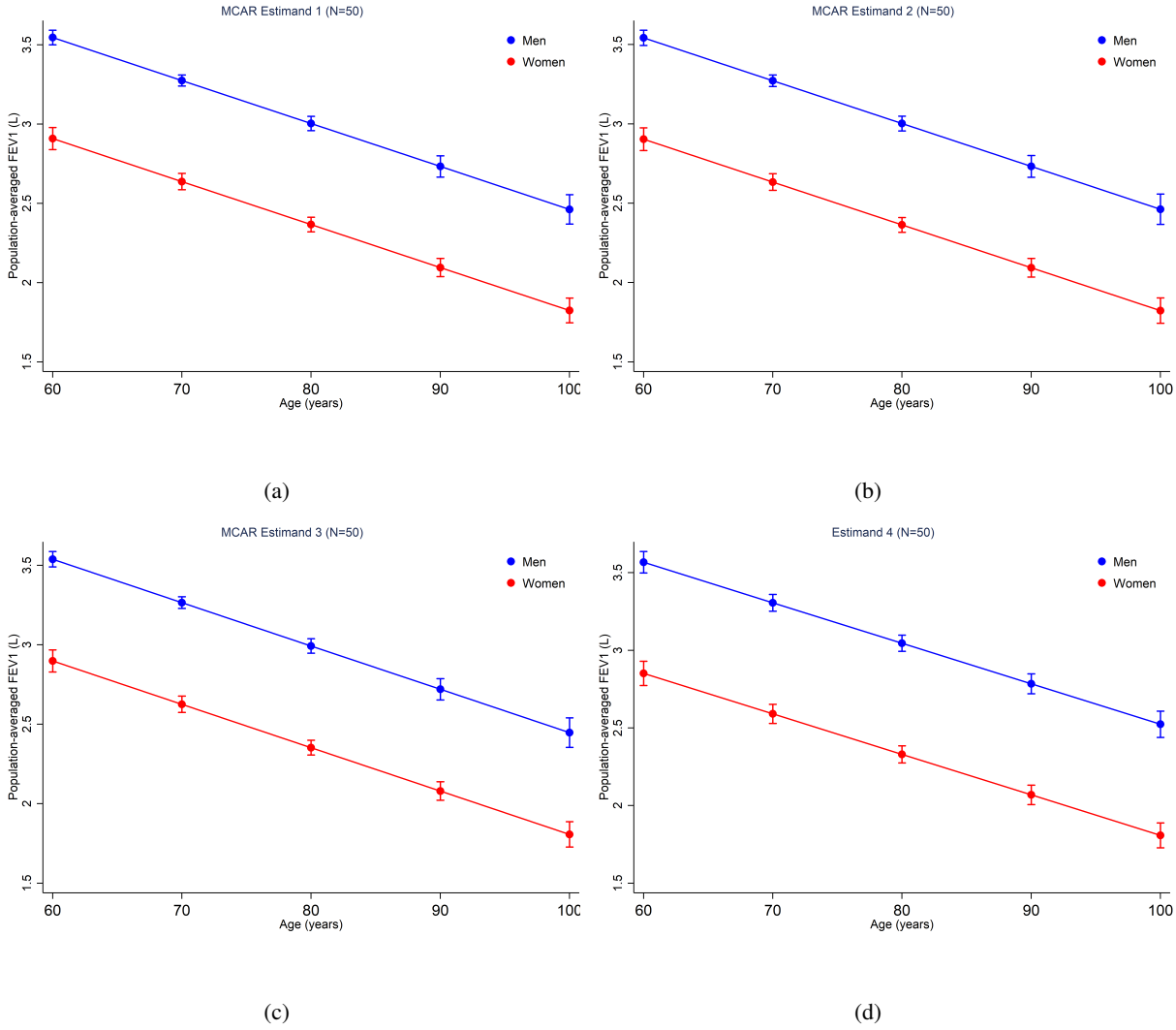
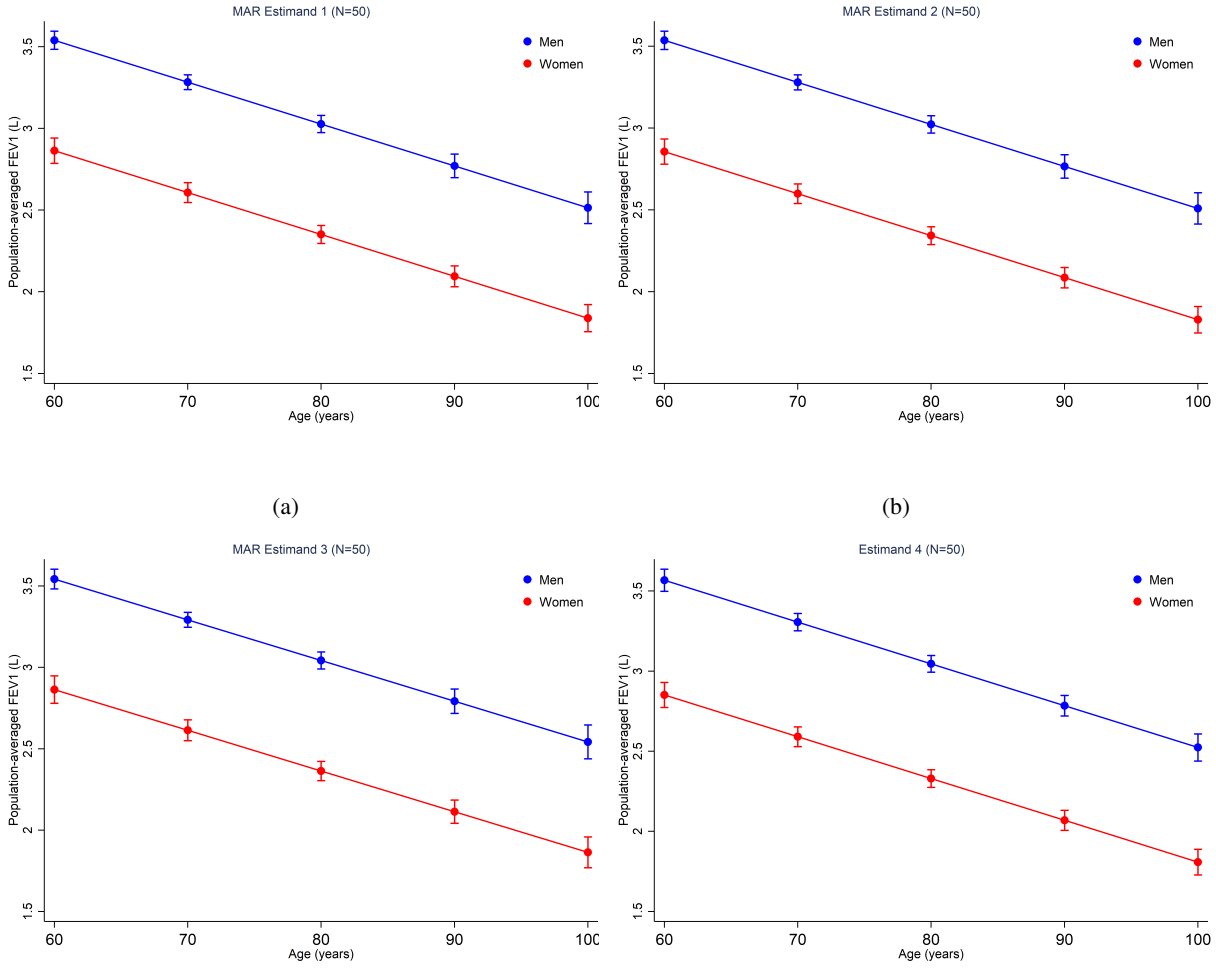


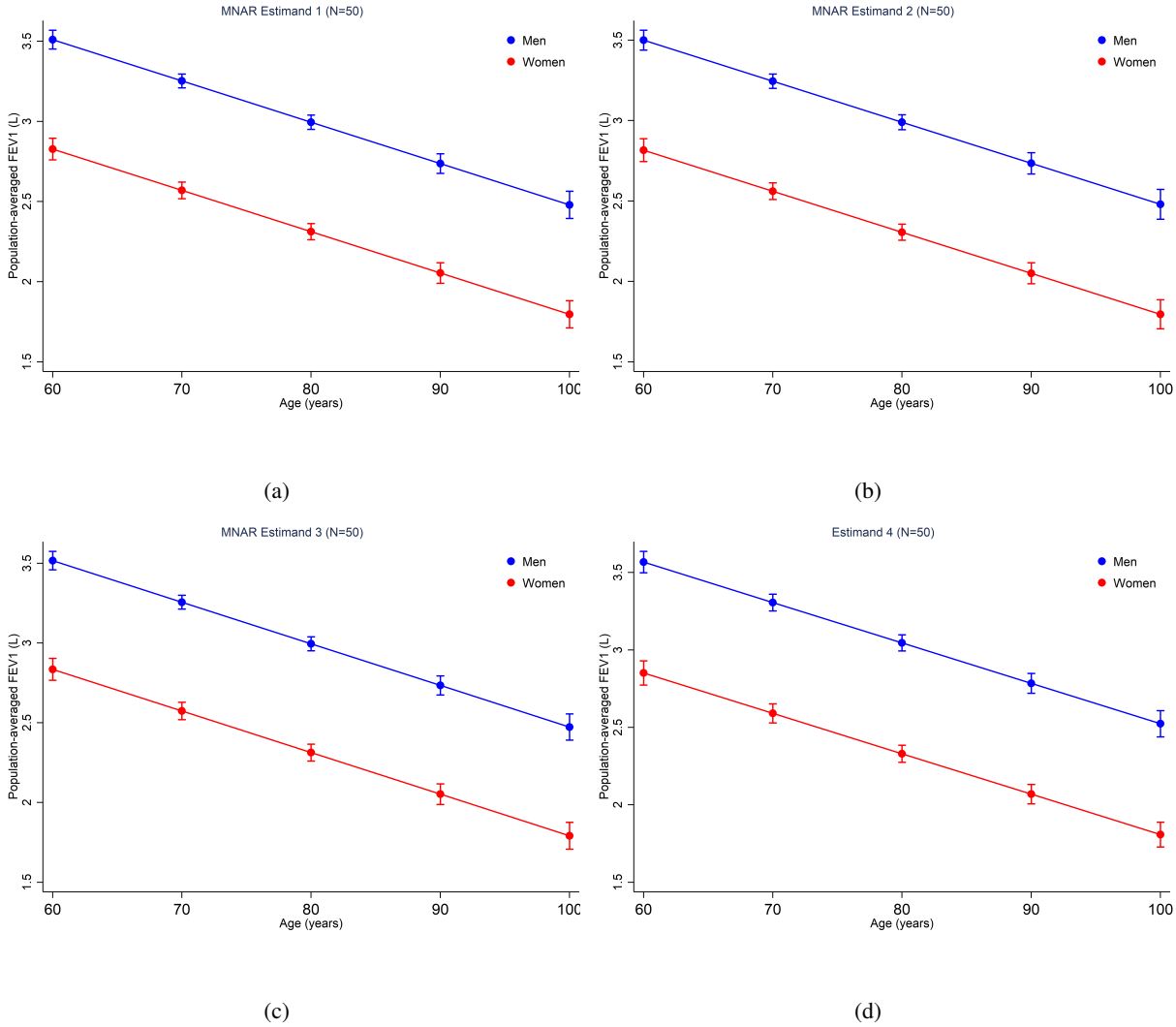
Figure 4.2: Predicted FEV1 values based on estimated estimands 1-3 under the MCAR mechanism and estimand 4, for different selected age groups, by sex, N=50.



(a) (b)

(c) (d)

Figure 4.3: Predicted FEV1 values based on estimated estimands 1-3 under the MAR mechanism and estimand 4, for different selected age groups, by sex, N=50.



(a) (b)

(c) (d)

Figure 4.4: Predicted FEV1 values based on estimated estimands 1-3 under the MNAR mechanism and estimand 4, for different selected age groups, by sex, N=50.

4.2 Large cohort

In table 4.3, the average weights are presented and in table 4.4 the largest weights are presented, for each estimand and visit. In comparison to the small cohort, the mean weights are even closer to 1 and the largest weights (around 3) are not as large as in the small cohort (around 6). The MAR estimands 1 and 2 provide both the smallest mean weights (0.91 for visit 5) and rather large maximum weight, for visits 3, 4 and 5 in the range 2.44-2.89. In addition, in this cohort the largest weights that stand out with respect to estimand and visit number are the MNAR estimands 1 and 2.

	Visit number (missing/dead proportions)			
Average weight per visit N=5000	**Visit 2 (35%/15%)**	**Visit 3 (500%/30%)**	**Visit 4 (70%/35 %)**	**Visit 5 (90%/45%)**
MAR - Estimand 1	1,00	1,00	1,00	1,07
MAR - Estimand 2	1,01	1,00	1,00	1,06
MAR - Estimand 3	1,00	1,00	1,03	0,93
MAR - Estimand 1	1,00	1,01	0,98	0,91
MAR - Estimand 2	1,01	1,01	0,98	0,91
MAR - Estimand 3	1,00	0,99	0,98	0,95
MAR - Estimand 1	1,00	1,00	1,00	0,94
MAR - Estimand 2	1,01	1,00	1,01	0,93
MAR - Estimand 3	0,99	0,99	0,96	0,94

Table 4.3: Average IP weights for all estimands, by visit, N=500.

	Visit number (missing/dead proportions)			
Average weight per visit	**Visit 2 (35%/15%)**	**Visit 3 (500%/30%)**	**Visit 4 (70%/35 %)**	**Visit 5 (90%/45%)**
MAR - Estimand 1	1,17	1,37	1,30	2,62
MAR - Estimand 2	1,500	1,51	1,65	2,49
MAR - Estimand 3	1,05	1,17	1,56	1,82
MAR - Estimand 1	1,87	2,44	2,48	2,43
MAR - Estimand 2	1,90	2,44	2,89	2,89
MAR - Estimand 3	1,66	1,63	1,74	1,77
MAR - Estimand 1	1,62	1,74	3,09	2,41
MAR - Estimand 2	2,00	1,74	3,57	2,45
MAR - Estimand 3	1,24	1,33	2,54	1,67

Table 4.4: Largest IP weights for all estimands, by visit, N=500.

In figure 4.5, the estimated age coefficients for all estimands are depicted, along with 95% confidence intervals. Comparable with the estimated age coefficients for the small cohort, the MCAR estimands provide smaller estimated age coefficients as compared to the other estimands, with the largest difference when compared to estimand 4. All estimates for estimands 1-3 are smaller than the estimate for estimand 4. The width of the confidence intervals for the age coefficients are roughly the same, approximately 0.0015 units.

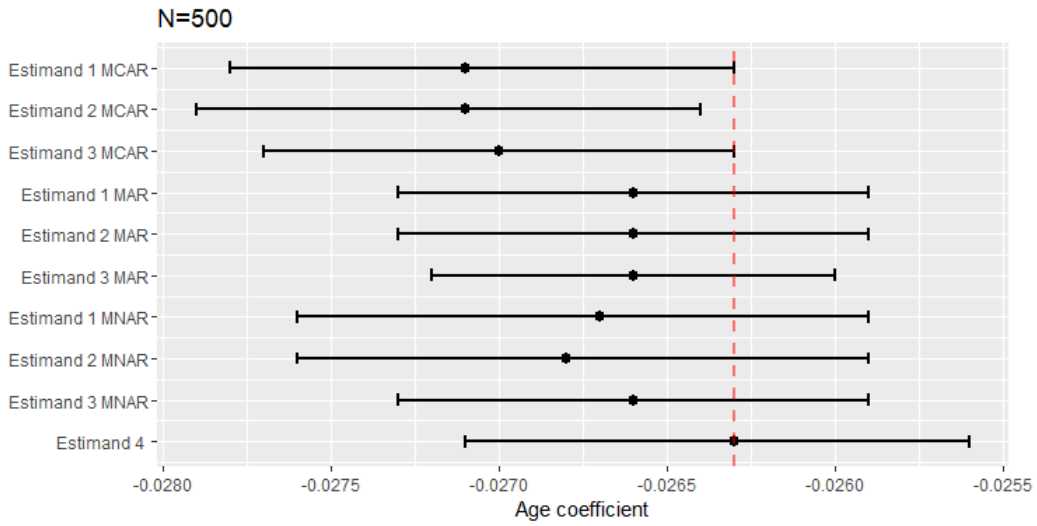


Figure 4.5: Estimated age coefficients from IPWGEE model, for all estimands, N=500.

The predicted population-averaged FEV1 for the large cohort are illustrated in figures 4.6, 4.7 and 4.8. All 95% confidence intervals are significantly narrower in comparison to the small cohort intervals, for all depicted ages and both sexes. This is likely due to the larger population and hence which imply smaller standard errors.

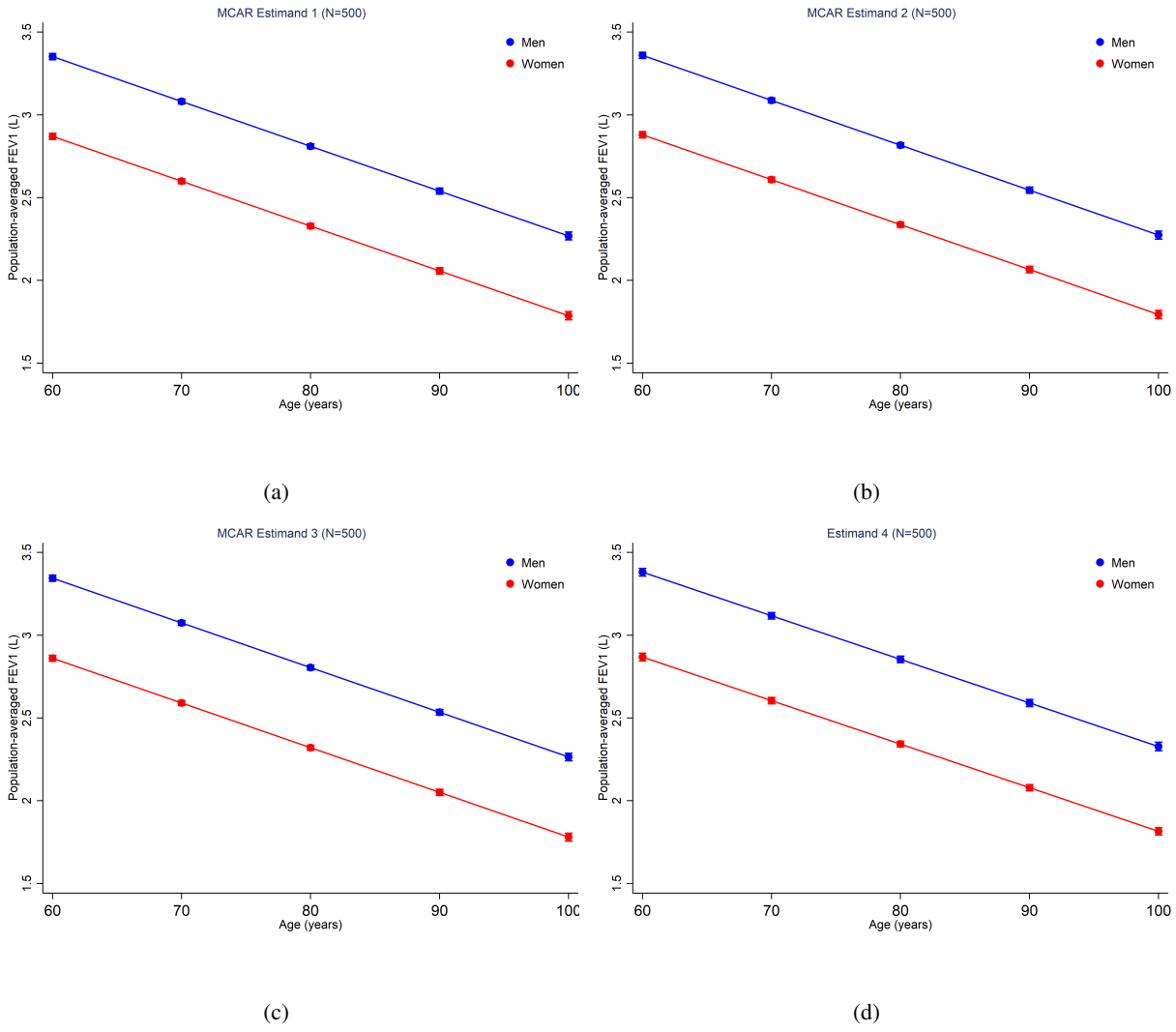
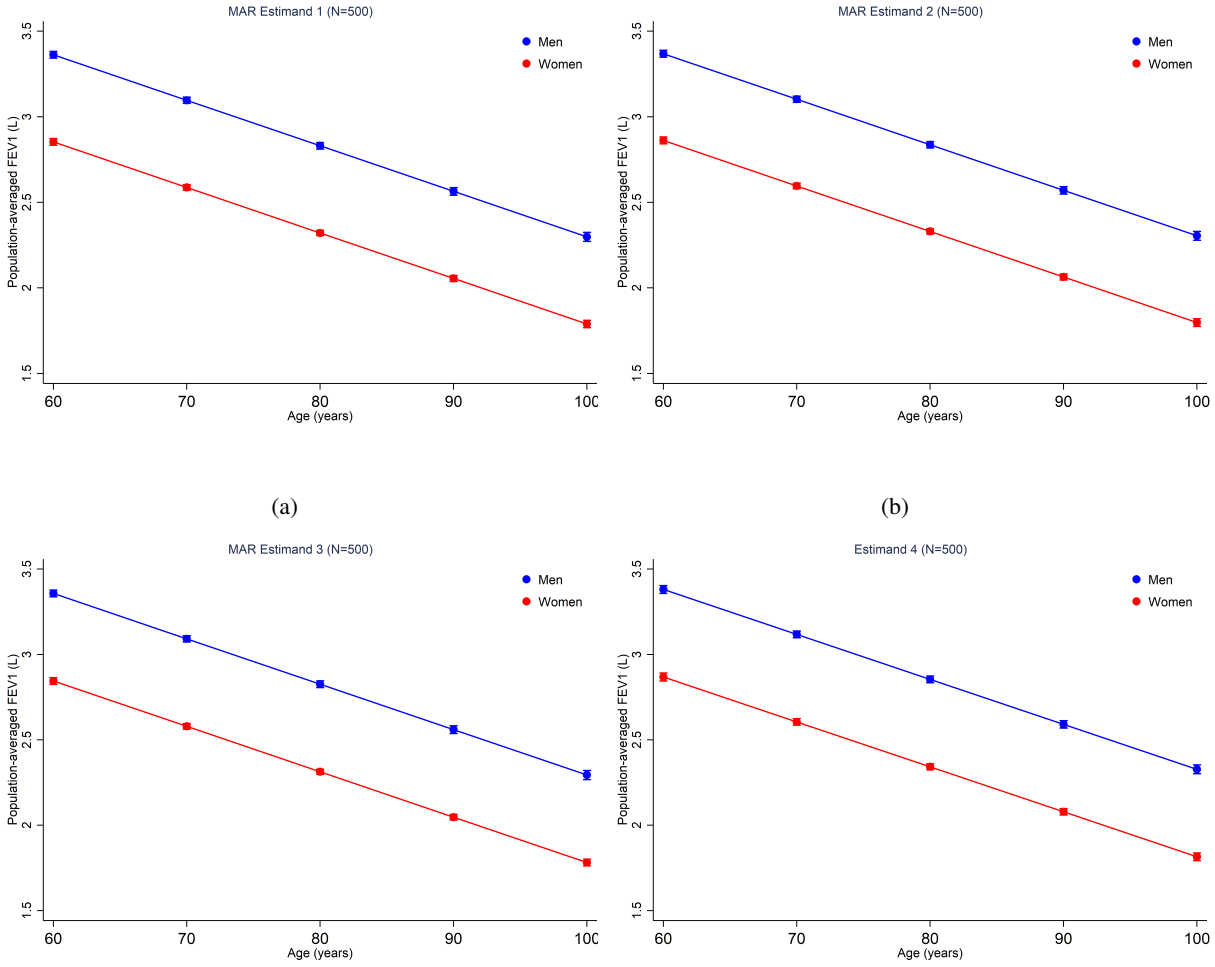


Figure 4.6: Predicted FEV1 values based on estimated estimands 1-3 under the MCAR mechanism and estimand 4, for different selected age groups, by sex, N=500.



(a) (b)

(c) (d)

Figure 4.7: Predicted FEV1 values based on estimated estimands 1-3 under the MAR mechanism and estimand 4, for different selected age groups, by sex, N=500.

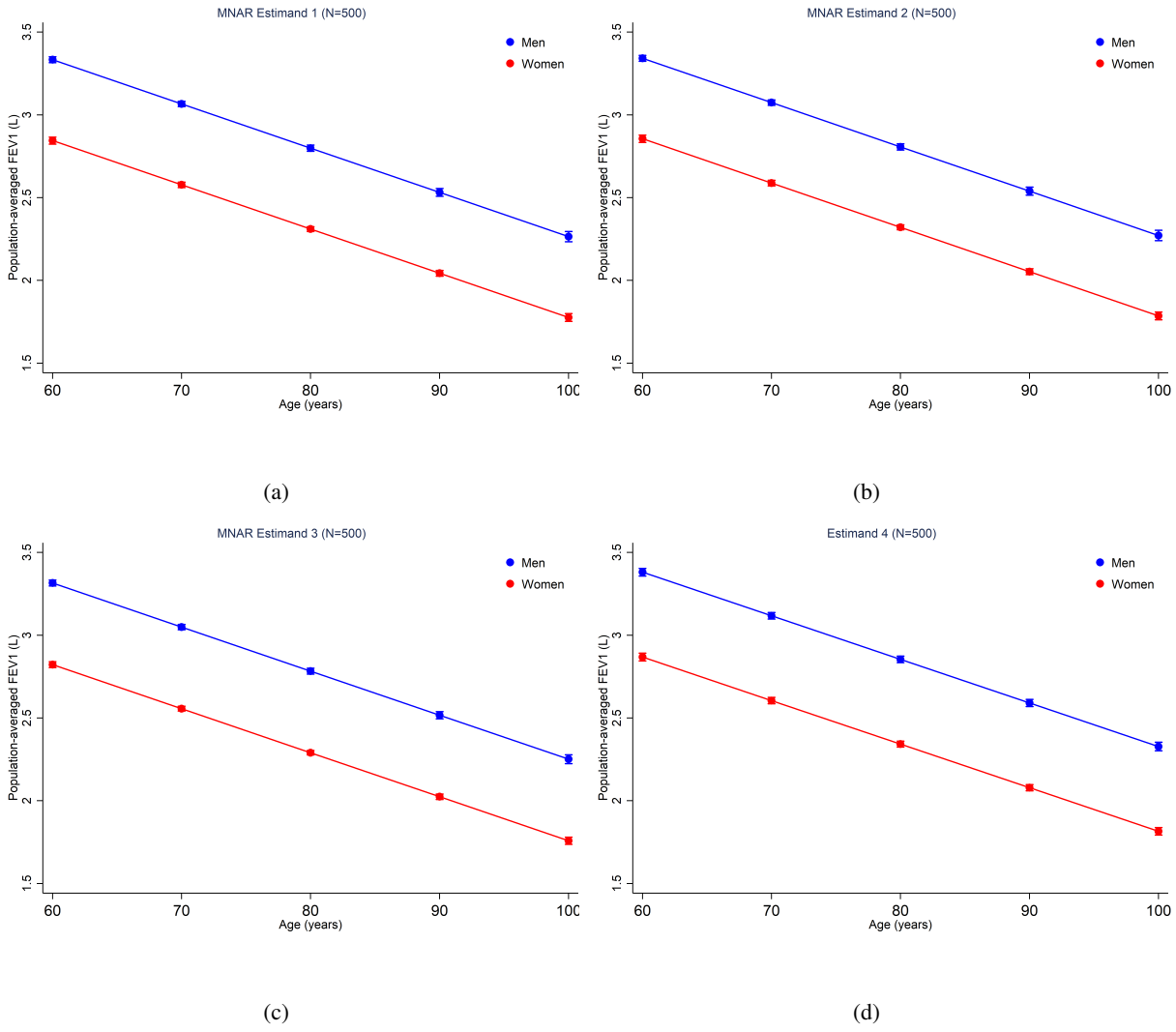


Figure 4.8: Predicted FEV1 values based on estimated estimands 1-3 under the MNAR mechanism and estimand 4, for different selected age groups, by sex, N=500.

Chapter 5

Discussion

In this thesis, four estimands with different strategies to account for missing data and distinguish reasons for missingness, by non-attendance or by death, in a longitudinal study investigating the lung function measured by FEV1 in an elderly population were studied. Three estimands used IPWGEE to model the missing data whereas a fourth estimand used unweighted GEE applied to data from alive participants only. In addition, three different missing mechanisms were explored: MCAR, MAR and MNAR.

Within each cohort size and per each age group and sex, all estimands predicts similar values, indicating that the method is not sensitive to whether an immortal (estimands 1 and 2) or mortal cohort (estimand 3) is used. Furthermore, the size of the IP weights are satisfactorily close to one on average, with few outliers, implying that the stabilised weights work well in this setting.

The results indicate that the differences between the models are rather driven by the underlying missingness mechanism, MCAR, MAR and MNAR, as opposed to the distinct handling of missing data by reason for missingness (estimands 1-3). The estimated age coefficients based on MCAR data show the largest bias in relation to estimand 4, as compared to the other missing mechanisms. On the other hand, estimand 4 exclude data from dead participants, and the dead mechanism is simulated to be MAR, depending on the observed FEV1 measurement from visit1, and this might be an explanation to why the results based on MAR data are less biased as compared to estimand 4. In addition, the MNAR data are simulated to be dependent on the unobserved FEV1 measurement at the missed visit. Since the FEV1 measurements are strongly correlated within study subjects, the dead MAR mechanism and the missing MNAR mechanism are strongly associated. This might also explain why the MNAR results are less biased than the MCAR data.

Since the underlying missingness mechanism is impossible to be verified and the possibility for MNAR data can never be ignored, expert knowledge of associated variables to specify the missingness model (IPW model) accompanying solid sensitivity and supplementary analyses by various assumptions are

essential for reliable inferences.

The inferences based on a mortal cohort provide estimates applicable for a population that are alive during the complete follow-up period. The participants that die at some time point during follow-up contribute to the models only until they die but are not accounted for in the missing data handling after death.

As opposed to the mortal cohort estimands, the immortal cohort estimands account for missing data by death and consequently make inferences based on data that are undefined.

The interpretability of the results depends on the chosen estimands since the estimands target different populations. In addition, one can question the relevance of an immortal cohort estimand since the missing data by death are undefined. Since the missing observations are undefined, the pseudo-population that is created by the IP weights has clear interpretation.

An important aspect to keep in mind when accounting for missing by death, is that frail participants can be assigned large weights to compensate for other frail participants that have died even if the proportion of frail patients are relatively small in the study population. For example, if lung function is the outcome of interest and IP weighting result in an over-representation of frail participants, the results might suggest that the estimated population-averaged lung function is worse than it is.

The opinions on when to use an immortal cohort diverge in the literature. Dufouil et al (2004) [41] argue that mortal cohort estimands are preferred unless the missing data by death is MCAR, in which case an immortal cohort can be used. In contrary Aalen et al. 2010 [42] claim that, depending on the study objective, an immortal cohort analysis can provide more clinically relevant results. They reason that even if the mortal cohort estimands seem to provide more relevant results, estimation based on an immortal cohort can take precedence when, for example, comparing two treatments and the proportions of dead patients are unequal between the treatment groups. Suppose that the outcome of interest is lung function measured by FEV1. The lung function is expected to decline by age. If one of the treatment groups have a longer survival time for patients as compared to the other group, patients in that group will have a lower number of deaths as compared to the other treatment group during the follow-up period. Then, limiting the analysis to observations from surviving patients might show that the treatment group with a smaller amount of death, where the patients live longer, have a lower lung function just because the patients live longer. For the purpose of comparing treatments, an immortal cohort, adjusting for death, might give more applicable insights.

IPW and MI are two distinct methods for missing data handling. While IPW requires a correct specification of the model for the probability of an observation to be missing, MI requires a specification of

the distribution of missing data, given the observed variables, i.e. an imputation model. Seaman et al. 2013 discuss the pros and cons of using IPW as compared to MI. They argue that an advantage of the MI approach, given a correctly specified imputation model, is that it is more efficient than the IPW approach given a correctly specified missingness model. A reason for this is that one assumes a the distribution of the missing data and implicitly impute data even if variables that predict missingness are unobserved. This means that, if the assumption of the underlying distribution of missing data is correctly specified, more information can be used in the handling of missing data and provide greater efficiency. On the other hand, IPW has the advantage of being easier to implement and understand, which can be useful in some contexts, for instance when explaining an analysis to non-statisticians. Seaman et al. also argue that IPW is preferred for handling of MAR data since the missingness model is more likely to be correctly specified for IPW as compared to the imputation model for MI. However, since the MAR assumption is an untestable assumption, it is impossible to know if the MAR assumption hold [33].

In addition, an advantage of MI is that it can handle intermittent missing data, whereas standard IPW cannot. Even if the simulated data included intermittent missing data patterns, the implementation of the computation of IP weights was implemented so that an observation could only be assigned a weight if the participant was observed at the last visit. However, Sun et al. (2018) proposed an approach to use IPW for handling of this intermittent missing data involving modelling of the missingness pattern incorporated into the estimation of IP weights [32].

5.1 Limitations and further research

There exist plenty of additional research questions related to the distinction in handling of missing data by non-attendance or by death, which may give additional insight. The following topics and alternative estimands are encouraged to be investigated further.

First, it is important to note that this simulated data set lacks many aspects of RWD, such as unknown underlying missing mechanisms, less quality of data and reduces capability to collect all relevant variables to name a few. Hence, simulated data tend to give more agreeable results as compared to RWD. In this study, it has been shown that possible differences between different estimand can be hard to detect. The first topic for further research is the application of IPWGEE to RWD. This has been investigated by Salazar et al (2016) [43].

In order to evaluate the estimands with the objective to analyse the lung function in an elderly population further, it would give supplementary insights to proceed from simulated data to the FEV1 data set from the GÅS study. Moving from simulated data to RWD anticipate added complexity to the estimation task. Not only because simulated data tend to be less entangled than RWD with respect to for example

associated variables, outliers and missing data, but also because the underlying missingness mechanisms will be unverifiable. On that account, this simulation study, in combination with the application on RWD, will increase the understanding of handling of missing data.

Another topic worth of attention is to investigate how different proportion of missingness reasons affect the results. In this thesis, the proportions for missing by non-attendance and by death were equal and kept fixed for all estimands. It is argued that the proportion of missing data by death should be considered when choosing the handling of missing data or intercurrent event suitable for the research objective.

Furthermore, a mixed model could be considered as opposed to the GEE approach if the objective would be to investigate individual participant's outcome trajectories. The mixed model can be combined with the IPW method in a similar way as the IPWGEE technique.

Moreover, finding the most suitable fitted model for missingness was not in the scope of this thesis and therefore the formulas 3.1, 3.2 and3.3 were kept constant for all estimands. However, since a correctly specified missingness model is a key requirement for IPW to provide valid inferences, it is of interest to test more combinations of formulas to ensure correct results, as suggested by Cole and Hernan (2008) [30].

Lastly, doubly robust estimators, where MI and IPW are combined, are suggested as an alternative that can give consistent results even if the missingness model is misspecified [21, 44]. Further investigation of this method as compared to the simple inverse probability weighting method used in this study, both in a simulation study and applied to RWD, is of interest.

Chapter 6

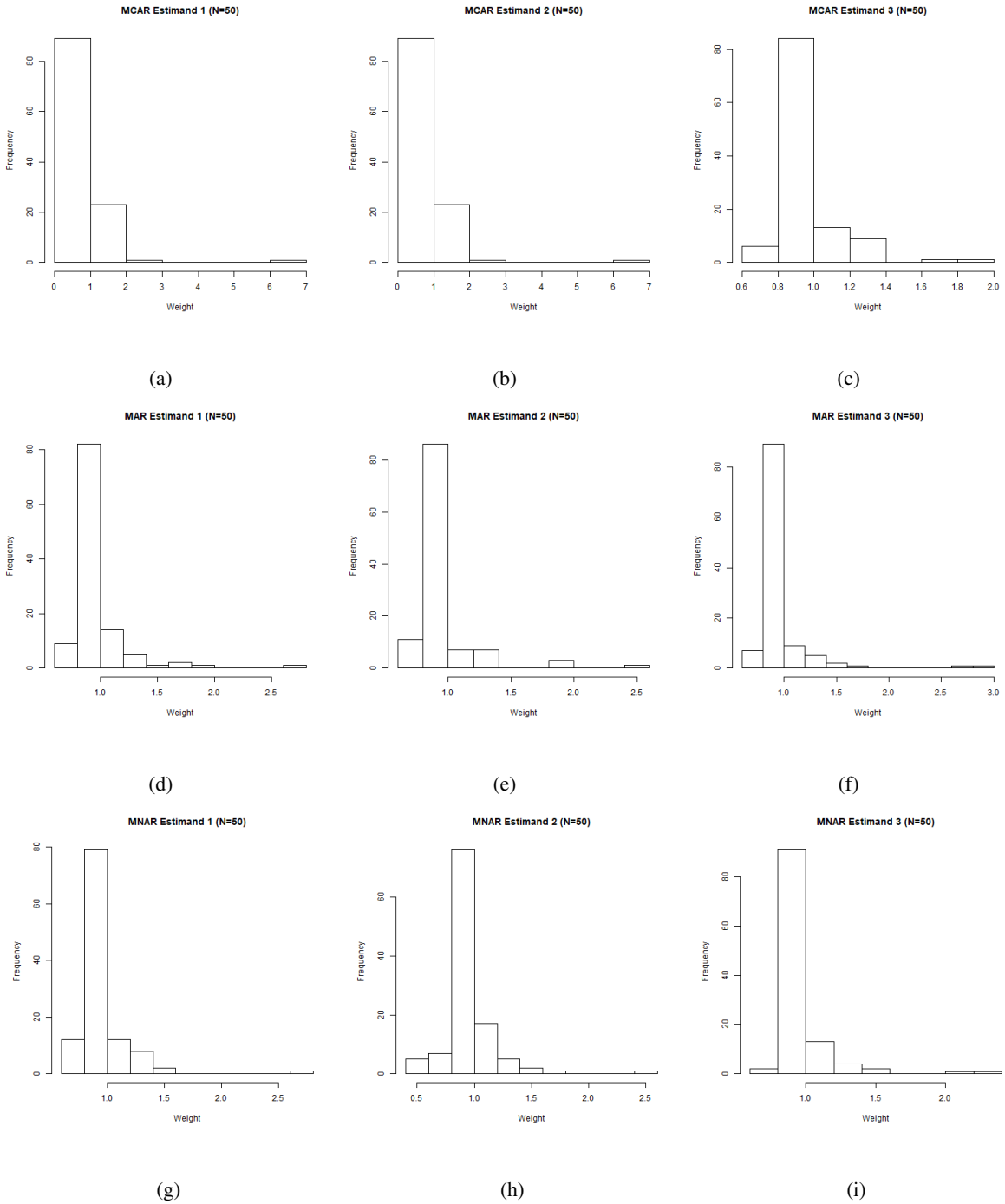
Conclusions

The estimation of all estimands resulted in similar predicted FEV1 values, based on a simulated data set. This suggests that the IPWGEE method is robust for different targets of inference. However, supplementary analyses and application on RWD will give further insight on possible differences between the estimands.

Through this project, it was highlighted that it is important to understand the implications on using estimands based on a mortal or immortal cohort. Whether a distinction of the handling of missing data by cause, for example by non-attendance or missing by death, should be made depends on the research question.

Appendix A

Additional plots

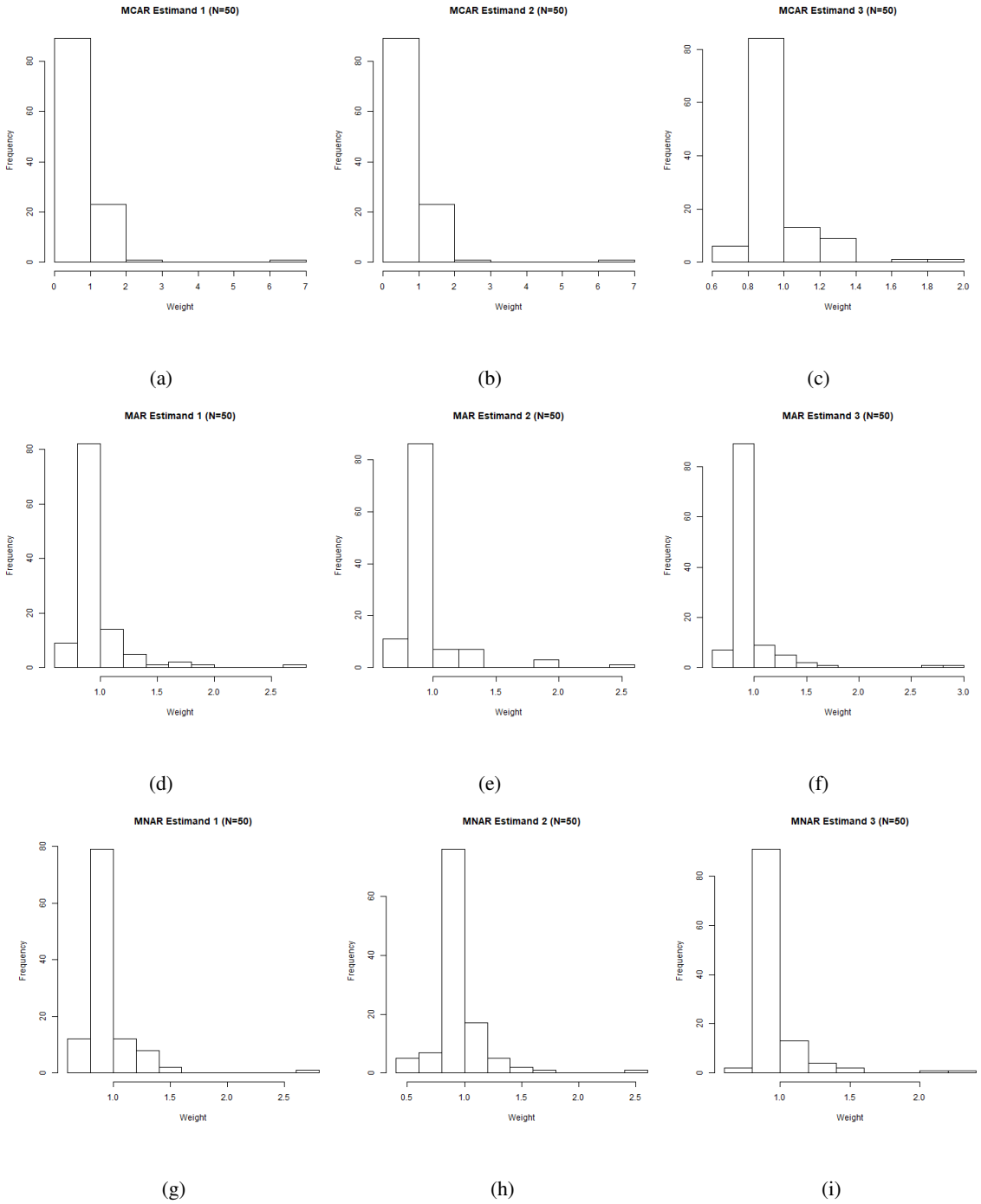


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure A.1: Histogram of IP weights for estimand 1-3, for all missing mechanisms MCAR, MAR and MNAR, for cohort of N=50.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure A.2: Histogram of IP weights for estimand 1-3, for all missing mechanisms MCAR, MAR and MNAR, for cohort of N=50.

Bibliography

[1] Jos W. R. Twisk. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN: 978-1-107-03003-9. DOI: 10.1017/CBO9781139342834.

[2] Alan Agresti. *An introduction to categorical data analysis*. Wiley series in probability and mathematical statistics. Hoboken, New Jersey, Chichester: Wiley, 2007. URL: http://dx.doi.org/10.1002/0470114754%20http://ludwig.lub.lu.se/login?url=https://ebookcentral.proquest.com/lib/lund/detail.action?docID=290465%20http://ludwig.lub.lu.se/login?url=https://onlinelibrary.wiley.com/doi/book/10.1002/0470114754.

[3] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied longitudinal analysis*. Hoboken, N.J.: Wiley, 2011. ISBN: 0-470-38027-6 978-0-470-38027-7.

[4] The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*. 2020. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf.

[5] Donald B. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592. ISSN: 0006-3444. DOI: 10.1093/biomet/63.3.581. URL: https://doi.org/10.1093/biomet/63.3.581 (visited on 05/15/2021).

[6] Karin Biering, Niels Henrik Hjollund, and Morten Frydenberg. "Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes". In: *Clinical epidemiology* 7 (2015). ISSN: 1179-1349. DOI: 10.2147/CLEP.S72247. PMID: 25653557. URL: https://pubmed.ncbi.nlm.nih.gov/25653557%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4303367/.

[7] Robins JM Hernán MA. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[8] B. F. Kurland et al. “Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims”. In: *Stat Sci* 24.2 (2009). Edition: 2010/02/02, p. 211. ISSN: 0883-4237 (Print) 0883-4237. DOI: 10.1214/09-sts293.

[9] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data”. In: *Journal of the American Statistical Association* 90.429 (1995), pp. 106–121. ISSN: 01621459. DOI: 10.2307/2291134. URL: http://www.jstor.org/stable/2291134 (visited on 05/29/2021).

[10] Robert E. Weiss. *Modeling Longitudinal Data*. Springer Texts in Statistics. New York, NY: Robert E. Weiss., 2005. ISBN: 978-0-387-28314-2. URL: http://dx.doi.org/10.1007/0-387-28314-5%20http://ludwig.lub.lu.se/login?url=https://link.springer.com/10.1007/0-387-28314-5.

[11] L. Wen, G. M. Terrera, and S. R. Seaman. “Methods for handling longitudinal outcome processes truncated by dropout and death”. In: *Biostatistics* 19.4 (Oct. 1, 2018). Edition: 2017/10/14, pp. 407–425. ISSN: 1465-4644 (Print) 1465-4644. DOI: 10.1093/biostatistics/kxx045.

[12] Publisher: Lund University. 2021. URL: http://www.geriatrik.lu.se/gott-aldrande-i-skane (visited on 05/30/2021).

[13] E. T. Thomas et al. “Rate of normal lung function decline in ageing adults: a systematic review of prospective cohort studies”. In: *BMJ Open* 9.6 (June 27, 2019). Edition: 2019/06/30, e028150. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2018-028150.

[14] Johannes Luoto et al. “Relative and absolute lung function change in a general population aged 60–102 years”. In: *European Respiratory Journal* (2018), p. 1701812.

[15] Place: Mayo Clinic. 2017. URL: https://www.mayoclinic.org/tests-procedures/spirometry/about/pac-20385201 (visited on 05/30/2021).

[16] Karen J. Tietze. “Chapter 5 - Review of Laboratory and Diagnostic Tests”. In: *Clinical Skills for Pharmacists (Third Edition)*. Ed. by Karen J. Tietze. Saint Louis: Mosby, 2012, pp. 86–122. ISBN: 978-0-323-07738-5. DOI: 10.1016/B978-0-323-07738-5.10005-5. URL: https://www.sciencedirect.com/science/article/pii/B9780323077385100055.

[17] Joseph M. Hilbe. “Generalized Linear Models”. In: *The American Statistician* 48.3 (1994), pp. 255–265. ISSN: 00031305. DOI: 10.2307/2684732. URL: http://www.jstor.org/stable/2684732 (visited on 06/07/2021).

[18] Raymond Myers et al. “Generalized Linear Models: With Applications in Engineering and the Sciences: Second Edition”. In: (). DOI: 10.1002/9780470556986.

[19] N. M. Laird and J. H. Ware. “Random-effects models for longitudinal data”. In: *Biometrics* 38.4 (Dec. 1982). Edition: 1982/12/01, pp. 963–74. ISSN: 0006-341X (Print) 0006-341x.

[20] Scott L. Zeger and Kung-Yee Liang. "Longitudinal Data Analysis for Discrete and Continuous Outcomes". In: *Biometrics* 42.1 (1986), pp. 121–130. ISSN: 0006341X, 15410420. DOI: 10.2307/2531248. URL: http://www.jstor.org/stable/2531248 (visited on 05/15/2021).

[21] Eric Vittinghoff et al. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Statistics for Biology and Health. Boston, MA: Springer US, 2012. URL: http://dx.doi.org/10.1007/978-1-4614-1353-0%20http://ludwig.lub.lu.se/login?url=https://link.springer.com/10.1007/978-1-4614-1353-0.

[22] James W. Hardin and Joseph M. Hilbe. *Generalized linear models and extensions*. College Station, Tex.: Stata, 2012. ISBN: 9781597181051 (pbk.) :

[23] Web Page. 2021. URL: https://online.stat.psu.edu/stat504/lesson/12.

[24] R. M. Daniel et al. "Using causal diagrams to guide analysis in missing data problems". In: *Stat Methods Med Res* 21.3 (June 2012). Edition: 2011/03/11, pp. 243–56. ISSN: 0962-2802. DOI: 10.1177/0962280210394469.

[25] Hyun Kang. "The prevention and handling of the missing data". In: *Korean journal of anesthesiology* 64.5 (2013). Edition: 2013/05/24, pp. 402–406. ISSN: 2005-6419 2005-7563. DOI: 10.4097/kjae.2013.64.5.402. PMID: 23741561. URL: https://pubmed.ncbi.nlm.nih.gov/23741561%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/.

[26] Lingling Li et al. "On weighting approaches for missing data". In: *Statistical methods in medical research* 22.1 (2013). Edition: 2011/06/24, pp. 14–30. ISSN: 1477-0334 0962-2802. DOI: 10.1177/0962280211403597. PMID: 21705435. URL: https://pubmed.ncbi.nlm.nih.gov/21705435%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998729/.

[27] "Encyclopedia of Research Design". In: (June 16, 2021). Number Of Volumes: 0. DOI: 10.4135/9781412961288. URL: https://methods.sagepub.com/reference/encyc-of-research-design.

[28] Magdalena Rosińska et al. "Potential adjustment methodology for missing data and reporting delay in the HIV Surveillance System, European Union/European Economic Area, 2015". In: *Eurosurveillance* 23 (2018). DOI: 10.2807/1560-7917.ES.2018.23.23.1700359.

[29] Paul R. Rosenbaum and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.41. URL: https://doi.org/10.1093/biomet/70.1.41 (visited on 05/23/2021).

[30] Stephen R. Cole and Miguel A. Hernán. "Constructing Inverse Probability Weights for Marginal Structural Models". In: *American Journal of Epidemiology* 168.6 (2008), pp. 656–664. ISSN: 0002-9262. DOI: 10.1093/aje/kwn164. URL: https://doi.org/10.1093/aje/kwn164 (visited on 05/23/2021).

[31] "BIOSTATISTICS WORKSHOP: MISSING DATA". In: (2016). Publisher: Sub-Saharan Africa CFAR meeting. URL: https://www.brown.edu/academics/medical/about-us/research/centers-institutes-and-programs/aids/sites/center-aids/files/Biostat_MissingData_LoriChibnik.pdf.

[32] B. Sun et al. "Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data". In: *Am J Epidemiol* 187.3 (Mar. 1, 2018). Edition: 2017/11/23, pp. 585–591. ISSN: 0002-9262 (Print) 0002-9262. DOI: 10.1093/aje/kwx350.

[33] S. R. Seaman and I. R. White. "Review of inverse probability weighting for dealing with missing data". In: *Stat Methods Med Res* 22.3 (June 2013). Edition: 2011/01/12, pp. 278–95. ISSN: 0962-2802. DOI: 10.1177/0962280210395740.

[34] Maria Iachina. "The Evaluation of the Performance of IPWGEE, a Simulation Study". In: *Communications in Statistics - Simulation and Computation* 38.6 (2009), pp. 1212–1227. ISSN: 0361-0918. DOI: 10.1080/03610910902859566. URL: https://doi.org/10.1080/03610910902859566.

[35] B. F. Kurland and P. J. Heagerty. "Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths". In: *Biostatistics* 6.2 (Apr. 2005). Edition: 2005/03/18, pp. 241–58. ISSN: 1465-4644 (Print) 1465-4644. DOI: 10.1093/biostatistics/kxi006.

[36] Ilya Lipkovich, Bohdana Ratitch, and Craig H. Mallinckrodt. "Causal Inference and Estimands in Clinical Trials". In: *Statistics in Biopharmaceutical Research* 12.1 (Jan. 2, 2020), pp. 54–67. ISSN: null. DOI: 10.1080/19466315.2019.1697739. URL: https://doi.org/10.1080/19466315.2019.1697739.

[37] C. J. Gore et al. "Spirometric standards for healthy adult lifetime nonsmokers in Australia". In: *European Respiratory Journal* 8.5 (1995), p. 773. URL: http://erj.ersjournals.com/content/8/5/773.abstract.

[38] Statistiska Centralbyrån (SCB). *Varannan svensk har övervikt eller fetma*. 2018. URL: https://www.scb.se/hitta-statistik/artiklar/2018/varannan-svensk-har-overvikt-eller-fetma/ (visited on 06/16/2021).

[39] James W. Hardin and Joseph M. Hilbe. *Generalized linear models and extensions*. College Station, Texas: Stata Press, 2018. ISBN: 978-1-59718-225-6 1-59718-225-7.

[40] R. Williams. "Using the margins command to estimate and interpret adjusted predictions and marginal effects". In: *Stata Journal* 12.2 (2012), pp. 308–331. URL: //.stata-journal.com/article.html?article=st0260.

[41] Carole Dufouil, Carol Brayne, and David Clayton. "Analysis of longitudinal studies with death and drop-out: a case study". In: *Statistics in Medicine* 23.14 (2004), pp. 2215–2226. ISSN: 0277-6715. DOI: https://doi.org/10.1002/sim.1821. URL: https://doi.org/10.1002/sim.1821.

[42] Odd O. Aalen and Nina Gunnes. "A dynamic approach for reconstructing missing longitudinal data using the linear increments model". In: *Biostatistics (Oxford, England)* 11.3 (2010), pp. 453–472. ISSN: 1468-4357 1465-4644. DOI: 10.1093/biostatistics/kxq014. URL: https://pubmed.ncbi.nlm.nih.gov/20388914%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3293429/.

[43] Alejandro Salazar et al. "Simple generalized estimating equations (GEEs) and weighted generalized estimating equations (WGEEs) in longitudinal studies with dropouts: guidelines and implementation in R". In: *Statistics in Medicine* 35.19 (Aug. 30, 2016), pp. 3424–3448. ISSN: 0277-6715. DOI: 10.1002/sim.6947. URL: https://doi.org/10.1002/sim.6947 (visited on 05/17/2021).

[44] D. Y. Kang Joseph and L. Schafer Joseph. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data". In: *Statistical Science* 22.4 (), pp. 523–539. DOI: 10.1214/07-STS227. URL: https://doi.org/10.1214/07-STS227.

Master's Theses in Mathematical Sciences 2021:E58
ISSN 1404-6342

LUNFMS-3103-2021

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/