Missing Data in the Swedish National Patients Register: Multiple Imputation by Fully Conditional Specification

Jesper Hörnblad
Missing Data in the Swedish National Patients Register: Multiple Imputation by Fully Conditional Specification.

Jesper Hörnblad

Master Thesis
Department of Statistics
Stockholm University
May 2013
Abstract

The national patient register includes information on discharged in- and outpatients from public and private care givers in Sweden. The case mix index, calculated from the diagnosis related group weights, may be used to compare the composition of patients in different hospitals. Missing information in the register, however, complicates the calculation of the case mix index. A multiple imputation model that is able to handle these missing observations would improve the knowledge about the produced health care, especially in hospitals with a lot of incomplete data. In this master thesis, complete case analysis, unconditional mean imputation as well as single and multiple imputation under a fully conditionally specified model were used to impute the diagnosis related group weights and calculate the case mix index. The analysis of their performance showed that all methods produced almost unbiased estimates as long as the data was missing completely at random. In contrast, when the missing data mechanism was depending on the value of the diagnosis related group weights, all methods produced biased results. Both, single and multiple imputation noticeably reduced the bias compared to complete case analysis and unconditional mean imputation. Further, even when the missing observations were missing completely at random, the complete case analysis yielded estimates with unnecessary wide confidence intervals reducing the power of statistical tests.

Keywords: bias, fully conditional specification, multiple imputation, single imputation.
Contents

1 Introduction 3

2 Missing Data 4
  2.1 Missing Data Mechanism ................................. 5
  2.2 Distinctness of Parameters ............................. 6
  2.3 Examine the Ignorable Missing Data Assumption ........ 6
  2.4 Missing Data Pattern ................................... 7

3 Handling Missing Data 8
  3.1 Multiple Imputation .................................... 10
  3.2 The Imputation Model .................................. 14
  3.3 Fully Conditional Specification .......................... 14
    3.3.1 Imputation Diagnostics .............................. 16
    3.3.2 MICE package in R ................................ 17

4 The National Patient Register 18
  4.1 Diagnos Related Groups ................................. 19

5 Results 20

6 Discussion 23

7 Conclusion 26

8 References 27

Appendix A: SAS and R codes 30
Appendix B: Imputation Diagnostics 33
1 Introduction

The national patient register (NPR) includes information on discharged in- and outpatients from public and private care givers in Sweden. The main purpose of the register is to observe the health care so that resources can be allocated in the best possible way and to ensure the quality of the health care. Diagnosis related groups (DRG) is a system used to classify patients into homogeneous groups according to diagnosis and resource utilisations. DRGs ability to describe the composition of patients in a hospital, i.e. the case mix, makes it useful for comparisons both over time and between different regions and care givers. To every DRG a weight is assigned. This weight is a relative measure of the average cost for a patient in a specific DRG compared to the average patient in the entire register. One way to compare the composition of patients in different hospitals is through the case mix index which is simply the average DRG weights in the specific hospitals (Socialstyrelsen, 2013a).

In 2011 about one and a half million inpatients were registered in the NPR. Among those, about one percent had missing information such that it was impossible to assign them to a DRG weight (Socialstyrelsens, 2013b). One percent may not seem to be much, however keeping the total number of patients in mind makes one to realise the impact this may have on the resources needed in the health care. Further, the amount of missing DRGs differs a lot between different county councils and different hospitals. For example, some hospitals have almost no missing observations while other have as much as 10 % missing values. The missing observations will therefore have greater impact on some hospitals.

There are many ways to handle incomplete data. Complete case analysis (CCA) is probably the most common. To discard all incomplete cases and rely on cases with complete records is simple but yields a loss of information, and relies on some naive assumptions to produce unbiased estimates. One alternative approach to the problem is to use imputation. By single imputation (SI) each missing observation is replaced by an imputed value to create a "complete" data set. By treating the imputed values as observed, uncertainty about parameter estimates will be underestimated, which may cause standard errors to be too small and p-values to be too low (Little and Rubin, 2002). Multiple imputation (MI), introduced by Rubin (1987), is a generally accepted method to handle incomplete data (Schafer and Graham, 2002). By imputing several plausible values for each missing observation, several "complete" data sets are created. By including the variability of the
multiply imputed data sets, imputation uncertainty is included in the standard errors. A suitable MI model would be able to improve the estimation of the case mix index, especially in hospitals with a lot of incomplete data.

The objective of this master thesis is to examine whether an appropriate MI model can be used to impute missing DRG weights for inpatients in the NPR.

The master thesis is organized as follows. In Section 1, the NPR is introduced and the problem caused by missing data in the register is briefly described. In Section 2, assumptions about missing data, including the missing data mechanism and the missing data pattern, are presented. In Section 3, some methods for handling missing data are described. Here, focus is on the MI and in particular, the fully conditional specification (FCS). In Section 4, the NPR is presented in more details and the DRG grouping mechanism is briefly described. In Section 5, MI is applied to one clinic in the NPR and the obtained results are compared to those obtained by CCA, mean imputation and SI regarding bias, coverage rates and widths of confidence intervals (CI). In the last section, the results and limitations of the master thesis are discussed.

2 Missing Data

Incomplete data may arise due to several different reasons including refusal, attrition, measurement errors or simply ignorance about of the individual asked question. No matter what the reason is, missing observations is a problem that has to be dealt with in all statistical areas (Allison, 2001). Besides making sure that the missing observations really are missing observations (Schafer and Graham, 2002), assumptions, explicitly or implicitly, about the missing data mechanism are always made. Assuming an ignorable missing data mechanism simplifies the analysis of the missing data as it means that the process causing the missing observations does not have to be explicitly modelled. The concept of ignorable missing data was introduced by Rubin (1976) as “the weakest simple conditions on the process that causes missing data such that it is always appropriate to ignore this process when making inferences about the distribution of the data”. For the missing data mechanism to be ignorable two conditions have to be fulfilled. First, the missing observations have to be missing at random (MAR). Second, the parameters in the missing data process have to be distinct from those in the the data.
The missing data pattern, describing which observations in the data that are missing, may further be useful to examine when dealing with incomplete data. A monotone missing data pattern (MMP) offers, for example, more flexibility in the choice of missing data method than an arbitrary missing data pattern (AMP) (Little and Rubin, 2002).

### 2.1 Missing Data Mechanism

The missing data mechanism describes the relationship between the missingness of the data and the values of the variables in the data matrix, i.e. whether the missingness depends on the underlying values of the variables in the data set. The missing data is usually said to be either missing completely at random (MCAR), MAR or missing not at random (MNAR).

Let \( Y = (Y^{obs}, Y^{mis}) \) be the complete data matrix consisting of the observed and unobserved values. Further, let \( R \) be a missing data indicator matrix, taking the value 1 if \( y_{ij} \) is missing and 0 if \( y_{ij} \) is observed.

The data is MCAR if the missingness is independent of any observation, observed or missing, in the data set:

\[
P(R|Y^{obs}, Y^{mis}) = P(R).
\] (1)

Thus, if the data is MCAR, the observed data can be seen as a random sub sample of the original sample and any estimate based on this sub sample may be considered to be unbiased.

Under the MAR assumption the missingness is allowed to depend on the observed data, \( Y^{obs} \), but not directly on the missing, \( Y^{mis} \). The data is MAR if:

\[
P(R|Y^{obs}, Y^{mis}) = P(R|Y^{obs}).
\] (2)

If the missing data cannot be assumed to be neither MCAR nor MAR, and the probability of \( R \) depends on the missing values even after taking the observed values into account,

\[
P(R|Y^{obs}, Y^{mis}) = P(R|Y^{obs}, Y^{mis}),
\] (3)

the data is said to be MNAR. In this case, the missing data mechanism cannot be ignored and has to be considered in the imputation model (Little and Rubin, 2002).
2.2 Distinctness of Parameters

Apart from assumptions about the missing data mechanism, assumptions also have to be made regarding the parameters of the missing data mechanism, $\xi$, in relation to those of the data, $\theta$. The distinctness of the parameters has a bit different meaning from the frequentist and the bayesian perspective. From the frequentist perspective, it means that the joint parameter space of $\theta$ and $\xi$ must be the product of the two individual parameter spaces. From the bayesian perspective, on the other hand, it means that any joint prior distribution applied to $(\theta, \xi)$ must factor into the independent marginal distributions for $\theta$ and $\xi$ (Schafer, 1997).

2.3 Examine the Ignorable Missing Data Assumption

Understanding why some observations are missing is crucial for the analysis of the incomplete data. At the same time, examining assumptions about the missing observations is inherently difficult since the missing observations are in fact unknown. Tests for the MCAR assumption have been suggested in the literature (see Little, 1988; Lee and Park, 1997 and Chen and Little, 1999), but there is no way to test if the data is MAR (Schafer and Graham, 2002).

There are some situations where the missing data is known to be at least MAR. These are mainly situations where the data is missing due to the study design and thus is under the control of the researcher, for example, with help of double sampling or randomized experiments with unbalanced design. As soon as the process leading to the missing data is out of the researchers control, one cannot be sure that the MAR assumption really holds (Schafer, 1997).

One way to increase the plausibility of the MAR assumption is to include auxiliary variables in the imputation process. Variables that are known to be highly correlated with the variables containing the missing observations can help to explain the missingness, thus making the MAR assumption more plausible. Auxiliary variables will also remove nonresponse bias that can be accounted for by the observed observations, reducing possible bias due to deviations from the MAR assumption (Collins, Schafer and Kam, 2001). Still, even though MAR is impossible to test for, it is the most commonly assumed missing data mechanism (Stuart et al., 2009).
2.4 Missing Data Pattern

The missing data pattern describes which values in the data matrix that are actually missing, and can help in the choice of method for handling the missing data. Missing data patterns are usually divided into monotone (MMP) and arbitrary missing patterns (AMP).

Let $Y_j$ be the $j$th variable, $j = 1, 2, ..., p$, in a data set. A MMP arises when the data can be ordered in such a way that having a missing value on variable $Y_j$ also means having missing values on all following variables $Y_{j+1}, ..., Y_p$ (see Figure 1). MMP often occurs due to attrition in longitudinal studies, where dropping out by definition means that all the following observations will be missing. A special case of MMP is the univariate missing data pattern (UMP) where only one variable in the data set suffer from missing observations, (see Figure 1). An AMP on the other hand arises when the data matrix cannot be ordered as in MMP, (see Figure 1). One example of AMP is item non response in surveys where respondents for some reason have failed to answer one or more questions, but missing values in one variable does not necessarily implies that all following variables are missing. (Little and Rubin, 2002)

$$A = \begin{bmatrix} x & x & x & x & x \\ x & x & x & ? & ? \\ x & x & ? & ? & ? \\ x & ? & ? & ? & ? \end{bmatrix}, \quad B = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & ? & ? \\ x & x & ? & ? \end{bmatrix}, \quad C = \begin{bmatrix} x & ? & x & x \\ x & x & ? & x \\ ? & x & ? & x \\ x & x & ? & ? \end{bmatrix}$$

Figure 1: Examples of missing data patterns. Rows correspond to units and columns to variables. Matrices A, B and C have MMP, UMP and AMP respectively.

MMP may simplify the analysis of the incomplete data as it may allow for the likelihood function to be factorized into factors for each block of cases with missing observations in the same variables, which can then be maximized separately. Methods constructed solely for MMP usually demand less computations than those designed also to handle AMP. It may sometimes even be worth considering removing a small number of observations or impute values for some variables using an arbitrary missing data method in order to
create a data set with a "monotone" missing data pattern (Little and Rubin, 2002).

In the next section some methods available to handle incomplete, relying on different assumptions about the missing data will presented.

3 Handling Missing Data

There are several different methods and strategies available to handle missing data. Graham (2009) defines three conditions that should be satisfied in a good method. First, the method should yield unbiased estimates of a variety of different parameters. Second, the method should include a way to assess the uncertainty about the parameter estimates, and third, the method should have good statistical power. It is further good to remember that the goal of any missing data procedure is not to recreate the missing observations but rather to retain the characteristics of the data and the associations between variables, such that valid and efficient inferences can be made (Schafer and Graham, 2002).

Complete (CCA) and available case analysis (ACA) are two types of methods that use case reduction to deal with the missing data. CCA is probably the most common way to handle incomplete data. In CCA the data set is edited to mimic a complete set by discarding all units containing any missing observations. While resorting to complete cases is simple, this method suffer from a loss of information in the incomplete cases and risk of bias if the missing data is not MCAR. In ACA all units with observed values on the variables of interest are used. This method makes better use of the data but still suffers from the loss of information. Further, ACA complicates the use of different statistics together, for example in tables, as the sample will differ for the different statistics. In ACA there is also a of risk producing correlations outside the natural bound [-1;1] (Little and Rubin, 2002).

Imputation on the other hand utilizes all collected information and allows the user to perform analysis of a "complete" data set using standard complete data methods. By single imputation (SI), each missing observation is filled in by one imputed value, creating one "complete" data set. SI methods range from ad-hoc methods like mean imputation to more complex regression imputations (Little and Rubin, 2002). If the aim of imputation is to recreate the missing observations as good as possible, imputing the conditional mean would probably be the best guess for every missing value. To
preserve associations between variables and provide valid estimates of the parameters, Little and Rubin (2002) however conclude that the imputations should be conditional on the observed data, multivariate and draws rather than means of the conditional distribution. Inefficiency and failure to incorporate imputation uncertainty in the standard errors are the two major drawbacks of SI. Failing to take into account the uncertainty caused by the fact that the imputed values are estimated from the the data may produce too small standard errors, narrow confidence intervals (CI) and low p-values (Little and Rubin, 2002).

According to Graham’s (2009) criteria, case deletion and SI can only be recommended in special cases. Case deletion will bias estimates of most parameters unless the missing observations are MCAR and have low power due to unnecessary wide CIs. SI may bias covariances and correlations, and at the same time underestimate variances and standard errors of the estimates.

Two methods that do meet Graham’s criteria are maximum likelihood (ML) and MI. These are further the generally recommended methods to handle incomplete data (Schafer and Graham, 2002). Under the MAR assumptions both methods yield consistent, asymptotically efficient and normally distributed estimates.

As with ordinary ML with complete data, the likelihood function, \( L = \prod_{i=1}^{n} f_i(y_{i1}, y_{i2}, ..., y_{ip}|\theta) \), is obtained and maximized with respect to \( \theta \). The likelihood with incomplete data is the product of the likelihood for all observations. The difference, in the case of incomplete data, is that the likelihood function is factorized into different parts according to the missing observations. If for example variables \( Y_1 \) and \( Y_2 \) contain missing observations but the rest are complete, the likelihood function would look like, \( L = \prod_{i=1}^{n} f_i(y_{i1}, y_{i2}, ..., y_{ip}|\theta) \prod_{i=1}^{m} f_i(y_{i3}, y_{i4}, ..., y_{ip}|\theta) \), where the second part is the marginal probability of the completely observed variables (Allison, 2012). This can of course be extended to include missing data on several variables by factorizing the likelihood into more than two parts. Among the different methods to maximize the likelihood function, the EM-algorithm (Dempster et al. 1977) is perhaps the most common.

MI is a general approach to deal with incomplete data. In contrast to SI, by MI several plausible values are imputed instead of just one for each missing observation. By imputing \( m \geq 2 \) random draws from a posterior distribution for every missing observation, \( m \) ”complete” data sets are created. Each data set is analyzed by a complete data method producing \( m \) point estimates. These estimates are then combined into one single estimate with
standard error consisting of both a within- and between-imputation variation component, properly reflecting the imputation uncertainty. By imputing several plausible values, the inefficiency problem in SI is resolved (Little and Rubin, 2002).

When comparing ML and MI, both their advantages and disadvantages should be considered. The greatest advantage of ML over MI is that ML is efficient while MI is only almost efficient (Allison, 2012). MI however has the great advantage that the imputations and the analysis can be done separately without putting the burden of dealing with the incomplete data on the researcher. In ML, handling the missing observations and performing the analysis have to be done simultaneously, putting a strain on the researcher who may not be familiar with the ways of dealing with incomplete data. Further, once the imputed data sets are constructed by MI, various statistical analyses can be conducted using the multiply imputed data sets.

In the next section MI will be considered in more detail. How to combine the estimates from the imputed data sets into one by the rules of Rubin (1987) and how to construct the imputation model by using fully conditional specification (FCS) will be described.

### 3.1 Multiple Imputation

MI is generally an appropriate method for dealing with incomplete data. When it was first suggested, it was mainly thought of as a way to handle nonresponse in a complex survey context (Rubin, 1987). However, it did not take long for MI to become accepted and useful in other settings as well (Rubin, 1996). Today the application of MI has grown and includes many statistical areas like clinical trials (Wood et al, 2004), epidemiology (Klebanoff and Cole, 2008) and longitudinal studies (Spratt et al., 2010).

The general idea behind MI is really quite simple and straightforward. Instead of imputing one value for each missing observation, \( m \geq 2 \) plausible values are imputed creating \( m \) ”complete” data sets. The difference between the imputed data sets reflects the uncertainty caused by imputations. The MI procedure includes three steps. First, create the multiply imputed data sets, preferably by random draws from a posterior distribution. Second, analyze each data set separately. Third, combine the results into a single set of parameters, standard errors and test statistics using the rules given by Rubin (1987).

Let \( Q, W, B \) and \( T \) be the quantity of interest, within imputation vari-
ance, between imputation variance and total variance associated with $Q$, respectively. Further, $\hat{Q}_d$ and $\hat{W}_d$, $d = 1, 2, \ldots, m$, are the estimated quantity and the variance that quantitate in the $d$th imputed data set. The combined point estimate of the parameter of interest, $\hat{Q}$, is the average of the $m$ estimates from the different completed data sets:

$$\hat{Q} = \frac{1}{m} \sum_{d=1}^{m} \hat{Q}_d. \quad (4)$$

The variability associated with $Q$ consists of two parts. The first part is the the within imputation variance, i.e. the average of the calculated variances associated with $Q$ in each imputed data set:

$$\hat{W} = \frac{1}{m} \sum_{d=1}^{m} W_d. \quad (5)$$

The second part is the between imputation variance, which is the variance of the $m$ estimates of $Q$,

$$\hat{B} = \frac{1}{m-1} \sum_{d=1}^{m} (\hat{Q}_d - \hat{Q})^2. \quad (6)$$

The total variance is then a modified sum of the within and between imputation variances,

$$\hat{T} = \hat{W} + \left(1 + \frac{1}{m}\right) B, \quad (7)$$

where $(1 + \frac{1}{m})$ is a correction for the finite number of imputed data sets. Confidence intervals and test statistics are based on a $t$-distribution with

$$v = (m - 1) \left(1 + \frac{1}{m+1} \frac{\hat{W}}{B}\right)^2 \quad (8)$$

degrees of freedom. Confidence intervals associated with $Q$ are calculated as

$$\hat{Q}_m \pm t_v(\alpha/2)\sqrt{\hat{T}}, \quad (9)$$

and the test statistic for testing hypothesis concerning $Q$ is

$$\frac{\hat{Q} - Q}{\sqrt{\hat{T}}} \sim t_v. \quad (10)$$

11
Even though the idea behind MI relies on a bayesian foundation, Rubin (1987) showed that under some general conditions MI will also yield valid inferences from a frequentist perspective. The conditions which leads to proper imputation are the following:

- As \( m \) becomes large, \((\hat{Q} - Q)/\sqrt{T}\) should approximately follow a standard normal distribution, \(N(0,1)\), with \( R \) and \( Y \) regarded as fixed.

- As \( m \) becomes large, \( \hat{W} \) should be a consistent estimate of \( W \) with random \( R \) and \( Y \) fixed.

- The between-imputation variance, \( \hat{B} \), should be stable over the repeated samples and lower then that of \( Q \).

As in the case with missing data mechanism it is usually difficult to determine whether the multiple imputations are actually proper or not. Schafer (1997) however states that from a practical viewpoint it is more important to examine whether the imputation method behaves well over repeated samples than if it actually, from a technical point of view, is strictly proper. This can be straight forward examined through simulations studies with realistic complete data populations.

Two main concerns regarding MI are discussed in the literature, the number of imputed data sets to yield valid inferences and the potential conflict between the model used for the imputations and that used for the analysis.

The number of imputations needed is a multifaceted question. MI is asymptotically efficient and for MI to be fully efficient infinitely many data sets would be needed (Allison, 2012). High efficiency may, however, be achieved with a quite low number of data sets. Rubin (1987) showed that as few as five data sets will lead to more than 90% efficiency. If statistical power is the main concern, more imputations may be needed. With 50% missing information, as many as 40 imputations may be needed for relatively high statistical power (Graham et al., 2007). Further, the estimated values of \( Q \) are draws from the posterior distribution \( p(Q|Y^{obs}, Y^{mis}, R) \). The inferences, like computation of CIs, based on the empirical distribution would demand thousands of draws, which in the case of MI would mean thousand of multiply imputed data sets (Rubin and Little, 2002). One of the features that make MI such a useful tool for handling incomplete data is the simplicity with which the analyst may analyze just a few imputed data sets and combine the obtained results into one. To store and analyze thousands of data
sets would greatly diminish this advantage. Rubin (1987), however, showed that if the complete-data posterior is based on a multivariate $t$-distribution, a very small number of imputed data sets is needed if the amount of missing data is not too large.

Regarding the potential conflict between imputation and analysis model, the concern has been that analyzing the data under a different model than the one used for the imputations would yield biased estimates. Partly, this is true, in particular if the imputation model is more restrictive than the analysis model. The general recommendation is that the imputation model should be at least as general as the analysis model (Schafer, 2003). Collins et al. (2001) concluded that an inclusive approach, including more variables and interactions, was greatly preferred to an restrictive approach when constructing the imputation model. Including auxiliary variables in the model both increased efficiency and reduced bias in the estimates. Including variables that may help explaining the missingness in the data further helps by increase the plausibility that the MAR assumption will hold (Schafer, 1997). Interactions should also be included in the model. Interactions left out in the imputation process will bias the the effect of the interactions and correlations between the variables towards zero (Graham, 2009).

The possibility to split the analysis of the incomplete data into the imputation and analysis part may, however, be an advantage rather than a disadvantage. By splitting the analysis into two steps, the data base administrator is able to use his/her special knowledge about the data set when the imputations are constructed. Confidential data, not available to the analyst, may for example be used as auxiliary variables to improve the imputations (Heitjan and Little, 1991). Even though it is recommended to make the imputation as general as possible, the communication between the person who conducted the imputations and the analyst is important. Clear documentation on how the imputations were created, problems in the imputation process and how to analyze the data sets should be provided together with the imputed data sets (Stuart et al., 2009).

The quality of the imputations rely almost entirely on the quality of the imputation model (Little and Rubin, 2002). There are several ways to construct the imputation model. In the next section, constructing the imputation model by a fully conditional specification (FCS) will be described.
3.2 The Imputation Model

The key step of the MI procedure is the specification of the imputation model, which is also the most difficult step (Little and Rubin, 2002). The imputation model is usually constructed through a joint model or a FCS (van Buuren et al., 2006). In joint modelling, the data is assumed to follow a known joint distribution, e.g. the multivariate normal, and imputations are then performed according to this joint model. Schafer (1997) suggested methods for imputing values under the multivariate normal, the loglinear and the general location model for continuous, categorical and mixed continuous and categorical variables, respectively. A joint model is however often difficult to specify. In this case, FCS may serve as an alternative. In the FCS approach, an imputation model, given all other variables is specified for each variable.

3.3 Fully Conditional Specification

In practice, the data often consists of variables with different scales, and quite complex relations between variables may occur that are hard to capture in an explicitly specified joint distribution for the entire data. By implementing MI under a FCS, a multivariate distribution is assumed, but specifying the explicit form of the joint model is unnecessary. Instead of drawing the imputations from a predefined joint distribution, imputations are generated sequentially variable-by-variable by specifying an imputation model for each variable given the other variables, \( P(Y_j^{mis}|Y_{-j}^{obs}, R) \). This even makes it possible to specify models for which no known joint distribution exist (van Buuren, 2007).

Let \( Y \) be the partially observed complete sample from the multivariate distribution \( P(Y|\theta) \). Further, let \( Y_{-j} \) be all variables in the data except \( Y_j \), \( j = 1, \ldots, p \). It is assumed that the distribution is completely specified by the vector of unknown parameters, \( \theta \), consisting of parameters specific to the respective conditional distribution and are not necessarily the product of the factorization of a “true” joint distribution. The posterior distribution is obtained by iteratively sampling from the conditional marginal distributions,

\[
P(Y_1|Y_2, Y_3, \ldots, Y_k, \theta_1) \\
\vdots \\
P(Y_k|Y_1, Y_2, \ldots, Y_{k-1}, \theta_k),
\]
that are assumed to completely specify the joint distribution. FCS starts with an initial imputation and draws imputations by iterating over the conditional densities constantly filling in the current draws of every variable. The \( t \)th iteration is thus the \( t \)th draw from the Gibbs sampler (van Buuren, 2012) as following

\[
\begin{align*}
\theta_1^{* (t)} &\sim P(\theta_1|Y^{obs}_1, Y_2^{t-1}, \ldots, Y_p^{t-1}), \\
Y_1^{* (t)} &\sim P(Y_1|Y^{obs}_1, Y_2^{t-1}, \ldots, Y_p^{t-1}, \theta_1^{* (t)}), \\
&\vdots \\
\theta_p^{* (t)} &\sim P(\theta_p|Y^{obs}_p, Y_2^{t-1}, \ldots, Y_{p-1}^t), \\
Y_p^{* (t)} &\sim P(Y_p|Y^{obs}_p, Y_1^{t-1}, \ldots, Y_{p-1}^{t-1}, \theta_p^{* (t)}),
\end{align*}
\]

(12)

where \( Y_j^{(t)} = (Y_j^{obs}, Y_j^{* (t)}) \) is the imputed value for the variable \( j \) at the \( t \)th iteration. As the cycle reaches convergence, the current draws are taken as the first set of imputed values. The cycle is then repeated until the desired number of imputations have been achieved (van Buuren et al., 2006).

FCS has many practical advantages over JM. Dividing the multidimensional problem into several one dimensional problems allows for more flexible models than if a joint model would be used. The joint distributions available for MI are rather limited while there exist many univariate distributions that can be used for imputation purposes. Hence, bounds, constraints and interactions between variables that may be difficult to include as a part of a multivariate model, can be more easily incorporated. Further, generalizations to data with nonignorable missing data mechanisms might be easier. Finally, different imputation models specified for every variable is easier to communicate to the practitioner (van Buuren et al., 2006).

FCS, however, suffers from the lack of theoretical justification. The conditional distributions, \( p(Y_1), \ldots, p(Y_j) \), are said to be compatible if the joint distribution, \( p(Y_1, \ldots, Y_j) \), factorizes into the \( j \) parts \( p(Y_j|Y_{-j}) \) for all \( Y_j \) so that the joint distribution exists and is unique. Incompatibility of the conditional distributions may be a problem. Convergence, and the distribution to which the conditionals converge, may or may not depend on the order of sequence of variables. This lack of theoretical justification may further cause problems because of difficulties when examining the quality of the imputations as the joint distribution may or may not exist, and convergence criteria
are ambiguous (van Buuren et al. 2006). Simulation studies have however showed the ability of FCS to produce unbiased estimates with confidence intervals with appropriate coverage, even under strong incompatible models (van Buuren et al, 2006; van Buuren 2007; Lee and Carlin 2010).

3.3.1 Imputation Diagnostics

The quality of the imputations should be examined. One way to do this is to assess whether the imputations are plausible or not, i.e. the imputed values should have been able to have been observed if the observations would not have been missing (van Buuren and Grothuis-Oudshorn, 2011). This may sometimes be really easy, for example, a man being admitted for problems during pregnancy would be a clear sign that something is wrong with the imputation model. Plotting the imputed values together with the observed and comparing the distributions of the imputed and the observed values is one way to detect possible implausible values. Abayomi et al. (2008) emphasize however that different distributions between the observed and imputed values is not sufficient to conclude that something is wrong. This is mainly due to the possibility that the missing observations are MNAR, but even if the missing observations are MAR somewhat different distributions may occur. Large differences is however a sign that this is something to look into. Outside subject specific knowledge should therefore also be used to check the plausibility of the imputations.

Another way to examine the quality of the imputations is to check the fit of the imputation model to the data where a bad fit is a sign of poor quality (Abayomi et al., 2008).

When the multiple imputations are conducted under a FCS, convergence of the iterations should be examined. The Gibbs sampler usually converge quite fast and usually no more than ten or twenty iterations are needed in each cycle. One way to assure convergence is to use many iterations. More iterations yield more intensive computations, and one should try not to use unnecessary many iterations. Plotting some parameter estimates against the number of iterations is the usual way to check the convergence (van Buuren and Grothuis-Oudshorn, 2011).
3.3.2 MICE package in R

There are a number of softwares that offers MI under a FCS including SAS, Stata, SPSS and R. In R the MICE package (van Buuren and Groothuis-Oudshorn, 2011) offers to perform the multiply imputations, analyze the imputed data sets and pool the results of the analysis under a FCS. MICE begins with an incomplete data frame and then stores the results from the three steps of the MI procedure in the mids, mira and mipo classes for the imputed data sets, estimated parameters and pooled results, respectively. MICE distinguishes between four kinds of variables: numeric, binary, nominal and ordinal, and provides a range of models to impute these. Apart from these ordinary imputation models, MICE also allows the user to perform passive imputation. Sometimes a transformation of one or many variables in the data is desired. One may for example want to have a log transformation or calculate the BMI from the weight and length. To make sure that the log transformation is sustained throughout the data, MICE allows the user to impute the log of the original variable instead of any regular imputation model. The MICE package further provides a number of plots and graphics to assess convergence and examine the quality of the imputations.

The FCS algorithm in MICE is a bayesian Gibbs sampler that samples from the conditional distribution until finally it draws are made the joint distribution of the variables. The uncertainty about the imputations is captured by both drawing imputations and the parameters of the conditional imputation model. It starts with a random draw from the observed values and cycles through the conditional distributions until convergence, or as long as is desired. The Gibbs samplers are run in parallel and in the last iteration the imputed values are taken to create the m imputed data sets (van Buuren, 2012). The MICE algorithm looks as follows:

1. Specify the imputation model $P(Y_{mis}^j | Y_{obs}^j, Y_{-j}, R)$.
2. Fill the in initial imputation values, $Y_j^0$, by random draws from $Y_j^{obs}$.
3. Repeat for iterations $t = 1, ..., T$.
4. Repeat for variables $j = 1, ..., p$.
5. Define $Y_{-j}^t = (Y_1^{t}, ..., Y_{j-1}^{t}, Y_{j+1}^{t}, ... Y_p^{t})$ as the data set currently completed for variables except $Y_j$. 

17
6. Draw $\theta_j^t \sim P(\theta_j^t | Y_j^{\text{obs}}, Y_{-j}^t)$.

7. Draw imputations $Y_j^t \sim P(Y_j^{\text{mis}} | Y_j^{\text{obs}}, Y_{-j}^t, R, \theta_j^t)$.

8. End repeat $j$.

9. End repeat $t$.

MICE is a very flexible package and allows the user to specify everything from the imputation models for each variable to the visiting sequence in the iterative procedure. Imputing categorical variables with many categories in MICE may cause problems as MICE creates dummy variables for every category potentially leading to very large matrices. Imputing categorical variables with many categories with a polynomous regression may thus be time consuming. In fact, trying to perform imputations on categorical variables with more than 50 categories will even produce an error massage. To speed up the calculations and be able to perform imputations on these kinds of variables van Buuren and Grothuis-Oudshorn (2011) recommend to impute categorical variables as numeric and use the predictive mean matching method (PMM) as soon as the number of categories is larger than 20. PMM only imputes values that have been observed and the original categories of the variable are thus preserved. This is however also the disadvantage of the package, as the main focus is on flexibility and not on resource minimization. Running MICE on big data sets may take time and demand a lot from the computer (van Buuren and Oudshoorn, 2011).

4 The National Patient Register

The National Patient Register (NPR) includes information collected each year by the Centre for Epidemiology (EpC), from the 21 county councils in Sweden (Socialstyrelsen, 2013b). For a complete description of the NPR, the reader is referred to socialstyrelsen.se. A short description will however follow to give the reader a picture of what the register is all about.

Information about both in- and outpatients is included in the register, but this master thesis will only consider inpatients. Information on 20 different variables is collected. A total of 1 598 864 observations were recorded during 2011. Of these, 16 206 had missing information in a way that made it impossible to assign them to a DRG weight. 6917 different main diagnoses,
148 hospitals and 69 different kinds of clinics were included in the register. The hospitals vary a lot both in size and characteristics of the patients. Some have more than 50,000 or even 100,000 visits during the year and others with less than 100 visits. The sorts of hospitals range from some only performing plastic surgery, to big hospitals treating all kinds of medical conditions. Further, the amount of incomplete data also differs a lot between the hospitals. Some hospitals have only complete records and some have very much missing data. In fact, 50% of the missing DRGs can be found in only 8 hospitals, and the hospital with most missing DRGs accounts for 15% of the total amount in the register. The main reason for missing DRGs are missing main diagnosis. However, a considerable amount of the missing DRGs are due to missing information about age.

One hospital in particular had a lot of incomplete data, and one specific clinic in that hospital accounted for almost all of this missing information. Of the 2775 patient visits the clinic, missing observations made it impossible to assign 2158 patients to a DRG weight. 2067 had missing diagnosis and 95 missing age.

4.1 Diagnos Related Groups

Diagnosis related groups (DRG) is a system used to classify patients into homogeneous groups according to diagnosis and resource utilisations. DRGs ability to describe the composition of patients in a hospital, i.e. case mix, makes it useful for comparisons both over time and between different regions and care givers (Socialstyrelsen, 2013a).

The algorithm for assigning patients to DRG is rather complicated and will not be described in detail here, instead a short description of algorithm will follow.

The process begins by grouping the main diagnoses into 26 major diagnostic categories (MDC) according to etiology and where the condition effects the body. Attached diagnoses may further affect the choice of DRG. If the attached diagnoses greatly affect the amount of resources needed, the patient is said to have a complicated version of the original DRG. Many, but not all, DRGs come in these CC pairs with one simple and one complicated version. The medical procedure performed may also influence what DRG the patient is assigned to. A surgical procedure is expensive and a patient undergoing surgery is assigned a different DRG than one that does not. Age may affect the choice of DRG in some cases. The most common limit is whether the
patient is below or above 18 years old. Other variables like sex and length of stay do not affect the choice of DRG, but may serve as an indicator of what main diagnosis, and thus DRG, is the most plausible when the imputations are being performed. Finally, for every case without enough information to be assigned to any particular DRG, a kind of "trash" DRG is assigned.

Every DRG is assigned a weight according to the average cost of that DRG in relation to the average cost of all patients. The average cost of all patients is assigned the point 1 and a higher point indicates a more expensive DRG. The DRG weight may be called a DRG point, and is then used to describe the amount of produced health care in a hospital or region. The average of the DRG point in a hospital is called the case mix index and may be used to compare the produced health care in different hospitals.

To examine the possibility to impute the DRG weights, imputations were performed on real data from an obstetrics and gynecology clinic in the NPR. Incomplete data was created in the data set with complete records. The quality of the imputations was assessed with respect to bias, coverage and widths of CIs. In the next section the results of this empirical study will be presented.

5 Results

One hospital, Hospital A, in the NPR had considerably more incomplete observations than the rest of the hospitals. In that hospital, the obstetrics and gynecology clinic was responsible for most of these incomplete observations.

To calculate bias, coverage rates and widths of CIs the true value of the case mix index is needed. An obstetric and gynecological clinic in another hospital, Hospital B, was located with almost no incomplete data, but similar to obstetric and gynecological clinic in Hospital A in other ways. The cases with missing observations where removed to create a complete data set with known true values of the DRG weights and the case mix index. To mimic the obstetric and gynecology clinic in Hospital A, incomplete data was created in the age variable and the main diagnoses. Any case with a missing age and/or a missing main diagnosis was also set to have a missing DRG weight. The amount of missing DRG weights was thus just under the sum of the missing age variables and the missing main diagnoses. Four different kinds of incomplete data were created. In the first data set 25% of the main diagnoses and 5% of the age values were MCAR. Two variables were simulated from
a random uniform distribution between zero and one, one to create missing observations among the main diagnoses, \( mhdia \), and one to create missing observations in the age variable, \( mage \). Every case with a value above 0.75 in \( mhdia \) got a missing value in the main diagnosis, and everyone with a value above 0.95 in \( mage \) got a missing value in the age variable. In the second incomplete data set, 50\% of the cases had a missing main diagnosis and 5\% had a missing age variable. The same method was used to create these missing observations, except that all cases with a value of \( mhdia \) above 0.50 got a missing main diagnosis. The missing observations were once again MCAR. To examine the sensitivity to a non random missing data mechanism, two data sets were created where the missing DRG weights depended on the DRG weight itself. The variable \( mhdia \) was used once again, but now the value in \( mhdia \) was multiplied with the DRG weight. Those with a low DRG weight thus had a greater chance of getting a low value on this variable. Those among the 25\% with the lowest score were given a missing value in the main diagnosis. The last data set was created in the same way, but now those among the 50\% with lowest scores got their main diagnosis deleted. The missing observations in the age variable were still 5\% and MCAR.

Unconditional mean imputation, SI and MI were performed and compared to the CCA. The single and multiple imputations were both conducted under the same FCS. To examine the performance of the methods, 1000 simulations were conducted and bias, coverage rates and widths of the CIs associated with the case mix index were calculated. Since the FCS is created through an iteration process the convergence must be assessed (van Buuren and Grothuis-Oudshorn, 2011). This was done by plotting the case mix index at each iteration. After 20 iterations, the case mix index and the standard errors had converged for all four incomplete data sets (see Figure 7, 8, 9 and 10 in Appendix B). In each simulation step, a FCS model was specified after 20 iterations and 10 imputed data sets were constructed. The variables used to perform the single and multiple imputations were: acute/planned admission, length of stay, age and whether a surgical procedure had been performed or not as they were those considered to have greatest effect on the DRG weights. The DRG weights were imputed using a bayesian linear regression (see Rubin, 1987). Age was measured in 5 year intervals and imputed using a polytomous logistic regression. The true value of the case mix index in the specified clinic was 0.7025.

With 25\% missing main diagnoses and 5\% missing ages all methods performed well with equally low almost ignorable bias. CCA had the widest CI
and mean imputation the narrowest (see Table 1).

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Bias</th>
<th>Coverage rate</th>
<th>Interval width</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.7116</td>
<td>0.0022 (0.313%)</td>
<td>*</td>
<td>0.0345</td>
</tr>
<tr>
<td>Mean imputation</td>
<td>0.7116</td>
<td>0.0022 (0.313%)</td>
<td>*</td>
<td>0.0247</td>
</tr>
<tr>
<td>SI FCS</td>
<td>0.7078</td>
<td>0.0053 (0.754%)</td>
<td>99.3%</td>
<td>0.0293</td>
</tr>
<tr>
<td>MI FCS</td>
<td>0.7077</td>
<td>0.0052 (0.740%)</td>
<td>100%</td>
<td>0.0325</td>
</tr>
</tbody>
</table>

* Coverage rates were not calculated for CCA and mean imputation.

Table 1: Evaluation of CCA, mean imputation and SI based on a FCS and MI based on a FCS with 25% of the main diagnosis being MCAR.

When the amount of missing observations increased, but still was MCAR all methods continued to produce almost unbiased estimates. The width of the CI increased for the CCA, but decreased for mean imputation and a little bit for SI. The decrease of the CI for the SI was enough for the coverage rate to fall the nominal value while MI still had 100% coverage (see Table 2).

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Bias</th>
<th>Coverage rate</th>
<th>Interval width</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.6981</td>
<td>-0.0044 (0.626%)</td>
<td>*</td>
<td>0.0405</td>
</tr>
<tr>
<td>Mean imputation</td>
<td>0.6981</td>
<td>-0.0044 (0.626%)</td>
<td>*</td>
<td>0.0189</td>
</tr>
<tr>
<td>SI FCS</td>
<td>0.6962</td>
<td>-0.0062 (0.882%)</td>
<td>91.6%</td>
<td>0.0276</td>
</tr>
<tr>
<td>MI FCS</td>
<td>0.6965</td>
<td>-0.0060 (0.854%)</td>
<td>100%</td>
<td>0.0363</td>
</tr>
</tbody>
</table>

* Coverage rates were not calculated for CCA and mean imputation.

Table 2: Evaluation of CCA, mean imputation and SI based on a FCS and MI based on a FCS with 50% of the main diagnosis being MCAR.

When the missingness depended on the DRG weights, the bias increased and coverage rates for both SI and MI fell well below the nominal level. There is however a significant difference in the bias for the CCA and mean imputation compared to SI and MI. In the worst case, with 50% of the main diagnoses being missing and the missingness depended on the DRG weights, all methods produced results that were seriously biased. Coverage rates were 0% for both SI and MI, which may be explained by the rather large bias (see Table 4). In both cases when the missing data mechanism was MNAR, the
bias for the SI and MI is about half of that of CCA and the unconditional mean imputation (see Table 3 and 4).

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Bias</th>
<th>Coverage rate</th>
<th>Interval width</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.7479</td>
<td>0.0453 (6.449%)</td>
<td>*</td>
<td>0.0360</td>
</tr>
<tr>
<td>Mean imputation</td>
<td>0.7478</td>
<td>0.0453 (6.449%)</td>
<td>*</td>
<td>0.0257</td>
</tr>
<tr>
<td>SI FCS</td>
<td>0.7252</td>
<td>0.0227 (3.231%)</td>
<td>1.8%</td>
<td>0.0299</td>
</tr>
<tr>
<td>MI FCS</td>
<td>0.7251</td>
<td>0.0226 (3.217%)</td>
<td>0%</td>
<td>0.0333</td>
</tr>
</tbody>
</table>

* Coverage rates were not calculated for CCA and mean imputation.

Table 3: Evaluation of CCA, mean imputation and SI based on a FCS and MI based on a FCS with 25% of the main diagnosis being MNAR.

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Bias</th>
<th>Coverage rate</th>
<th>Interval width</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.8130</td>
<td>0.1105 (15.730%)</td>
<td>*</td>
<td>0.0474</td>
</tr>
<tr>
<td>Mean imputation</td>
<td>0.8130</td>
<td>0.1105 (15.730%)</td>
<td>*</td>
<td>0.0227</td>
</tr>
<tr>
<td>SI FCS</td>
<td>0.7554</td>
<td>0.0529 (7.530%)</td>
<td>0%</td>
<td>0.0320</td>
</tr>
<tr>
<td>MI FCS</td>
<td>0.7552</td>
<td>0.0527 (7.502%)</td>
<td>0%</td>
<td>0.0412</td>
</tr>
</tbody>
</table>

* Coverage rates were not calculated for CCA and mean imputation.

Table 4: Evaluation of CCA, mean imputation and SI based on a FCS and MI based on a FCS with 50% of the main diagnosis being MNAR.

6 Discussion

The NPR contains information about discharged patients in private and public care givers in Sweden. The register is, among other things, used to compare the produced health care in different hospitals and county councils. The case mix index, calculated from DRG weights, may be used to compare the composition of patients between different care givers. Some patients are, however, missing information such that they cannot be assigned a DRG weight, this makes it difficult to calculate the case mix index.

This master thesis set out to examine if MI of the missing DRG weights could improve the estimation of the case mix index. CCA, unconditional
mean imputation, SI and MI were conducted on real data with four different kinds of missingness. SI and MI were conducted under the same FCS. Two data sets were created with DRG weights missing completely at random, one with 25% and with 50% missing DRG weights. Two other data sets were created, where the missingness of the DRG weights were made to depend on the weights themselves, one with 25% and with 50% missing observations. In addition, 5% of the cases, in each data set, had missing observations in the age variable. The performance of the methods was compared regarding bias, coverage rates and width of CIs associated with the case mix index.

All methods produced similar, almost unbiased, results when the missing observations were MCAR. Confidence intervals were the narrowest for mean imputation, followed by SI, MI and CCA. The coverage rates for MI were 100% both when the amount of missing DRG weights was 25% and 50%. In contrast, the coverage rates for SI fell below the nominal level of 95% when the amount of missing observations grew to 50%. These results were all to be expected, all methods should produce unbiased results of the mean when the missing observations are MCAR. Other parameters, would however be expected to be biased by mean imputation and SI. Further, SI treats the imputed values as true observed values. Failing to take the imputation uncertainty into account will in most cases lead to coverage rates below the nominal level (Little and Rubin, 2002).

When the missing data was MNAR, all methods produced biased results. This was also expected. CCA only yields unbiased estimates if the missing observations are MCAR and MI when the data is MAR (Little and Rubin, 2002). The coverage rates for SI and MI were well below the nominal level, 0% or almost 0%, probably due to the large biases of the estimates. One of the advantages of MI is that the variables in the imputation model may help to reduce the amount of the bias even when the missing data is MNAR (Schafer, 1997). In this study the variables in the imputation model managed to reduce the bias with just above 50%, in both cases, compared to the CCA and mean imputation.

One result that may seem strange is the coverage rates of SI and MI when the amount of missing main diagnoses was 25% and the missing data mechanism was MNAR. Even though the MI had wider confidence intervals than SI, SI still yielded a higher coverage rate, opposite to what would have been expected. This is most probably due to the fact that in SI relies on just one set of imputations while the point estimate from MI is the mean of ten imputed data sets. It is not unlikely that one imputed set is extreme and
manage to include the true estimate. If only one of the data sets in the MI is extreme, the other will however pull estimate of the case mix index towards the mean.

Graham (2009) established three criteria for a good missing data method. First, the method should yield unbiased estimates for a variety of parameters. Second, the method should include a way to assess uncertainty due to the missing observations. Third, the method should have good statistical power. Among the methods studied in this master thesis, MI was the only one that may claim to have fulfilled these criteria, when the missing observations were MCAR. MI had a coverage rate of 100% both with 25% and 50% missing DRG weights. SI fails to include the imputation uncertainty leading to coverage rates below the nominal level which was also the case in this master thesis. MI further produced narrower CIs than those from the CCA, leading to greater statistical power in hypothesis tests regarding the case mix index. To fully investigate the MI's ability to deal with the missing information, other parameters and associations between variables have to be investigated. When the missing data was MNAR, none of the methods can really be considered good. All of them produced severely biased estimates. The obtained results, however, indicate that MI at least would improve the estimation of the case mix index.

It is worth considering the limitation of this master thesis. First, only one clinic in one hospital was examined. Before imputations can be applied to the entire register further investigations would be needed. To begin with a subsample, is, however, a good way to examine the possibilities to perform MI, and detect possible problems the may impair the result of imputations (Stuart et al., 2009). One thing that should be taken under consideration is whether to use one imputation model for the entire register, or if the imputations should be performed under different models for each hospital and clinic. The data in the entire NPR would be best described by a multilevel model including the different levels of the register. Ignoring the distinct characteristics of different hospitals and clinics when constructing the imputation model would fail to preserve relationships and characteristics of the data set, which is the main goal of the imputations (Schafer and Graham, 2002). One way to handle this is to impute the missing observations under different models for different clusters (Graham, 2009), i.e. create different imputation models for the different hospitals and clinics. Further, the gains of a MI model should be weighed against the extra time and resources needed for the imputation process. Some hospitals with a lot of missing data would definitely benefit
from a well constructed imputation model, while other may do just as well by simply relying on the complete cases. There is no rule of thumb when to rely on complete cases and when to deal with the missing observations, rather this have to be considered for each case. Since a lot of the incomplete data is concentrated to a few hospitals, this would probably be a good solution. In cases where the amount of missing data is low these observations may just be ignored. Further, the very different characteristics regarding kinds of patients in different hospitals may even make one big imputation model undesirable.

Second, one limitation is caused by the software, i.e. the MICE package. Using this package does not allow the user to extract the imputation model. Checking the fit of the model to the data is one way to assess the quality of the imputations (Abayomi et al. 2009). The assessment of the imputations will thus solely rely on the comparisons of the observed and imputed values and external assessment of the plausibility of the imputed values.

7 Conclusion

Despite the limitations of this master thesis, and the biased estimates when the missing DRG weights were MNAR, the obtained results indicate that MI under a FCS may improve the estimates of the case mix index.

The missing data mechanisms studied in this master thesis were rather extreme, both regarding the amount of missing observations and the dependence on the DRG weights. Most clinics do not have as much missing main diagnoses as 50% or even 25%, and the missing data mechanism may sometimes properly be assumed to be MCAR. A less extreme missing data mechanism would reduce the bias of the estimates and improve the results of the MI. Further investigations are however needed before the method can be applied to handle the missing observations in the whole NPR.
8 References


Appendix A: SAS and R codes

*- Create data set with 25% MCAR DRG weights and 5% MCAR age -*;
data thesis.p2011slut_HospitalB4MCAR25;
set thesis.p2011slut_HospitalB4complete;
mhdia=ranuni(0);
mage=ranuni(0);
if mhdia>0.75 then hdia="";
if mhdia>0.75 then drg="";
if mage>0.95 then alder="";
if mage>0.95 then drg="";
drop mhdia mage;
run;

*- Create data set with 50% MCAR DRG weights and 5% MCAR age -*;
data thesis.p2011slut_HospitalB4MCAR50;
set thesis.p2011slut_HospitalB4complete;
mhdia=ranuni(0);
mage=ranuni(0);
if mhdia>0.50 then hdia="";
if mhdia>0.50 then drg="";
if mage>0.95 then alder="";
if mage>0.95 then drg="";
drop mhdia mage;
run;

*- Create data set with 25% MNAR DRG weights and 5% MCAR age -*;
data thesis.p2011slut_HospitalB4nyrg;
set thesis.p2011slut_HospitalB4nyrg;
mhdia=ranuni(0)*nyrg;
mage=ranuni(0);
run;
proc univariate data=thesis.p2011slut_HospitalB4nyrg;
var mhdia mage;
run;
data thesis.p2011slut_HospitalB4MNAR25;
set thesis.p2011slut_HospitalB4nyrg;
if mhdia<0.148187 then hdia="";
if mhdia<0.148187 then drg="";
if mage>0.95 then alder="";
drop mhdia mage nyrg;
run;
*- Create data set with 50% MNAR DRG weights and 5% MCAR age -*;
data thesis.p2011slut_HospitalB4MNAR50;
set thesis.p2011slut_HospitalB4nyrg;
if mhdia<0.29558 then hdia="";
if mhdia<0.29558 then drg="";
if mage>0.95 then alder="";
drop mhdia mage nyrg;
run;
# CREATE PREDICTOR MATRIX FOR MICE: 1 = USED AS PREDICTOR, 0 = NOT USED AS PREDICTOR #
pred.mat<-matrix(rep(0,36),nrow=6,ncol=6)
rownames(pred.mat)<-c("PVARD","VTID","age","surgery","nyrg","nyrgsq")
colnames(pred.mat)<-c("PVARD","VTID","age","surgery","nyrg","nyrgsq")
diag(pred.mat)<-0
pred.mat [c("age","nyrg"),"PVARD"] <-1
pred.mat [c("age","nyrg"),"VTID"] <-1
pred.mat [c("age","nyrg"),"age"] <-1
pred.mat [c("age","nyrg"),"surgery"] <-1
pred.mat ["nyrgsq","nyrg"] <-1
pred.mat [,"nyrgsq"] <-0
# ASSIGN IMPUTATION METHOD FOR EACH VARIABLE, "" = NO IMPUTATION #
method<-c(rep("" ,13))
method<-setNames(method,c("PVARD","VTID","age","surgery","nyrg","nyrgsq"))
method ["PVARD"] <-""
method ["VTID"] <-""
method ["age"] <-"polyreg"
method ["surgery"] <-""
method ["DRGweight"] <-"norm"
method ["nyrgsq"] <-" squeeze(nyrg,bounds=c(0,4))"
# SIMULATION WITH BIAS, COVERAGE AND INTERVAL WIDTH FOR hospital b #
test.impute <- function(data){
  imp.hospitalb<-mice(data, m=10, maxit=20, pred=pred.mat, meth=method, print=FALSE)
  fitMI <- with(imp.hospitalb, lm(nyrgsq 1))
}
estMI <- pool(fitMI)

tabMCAR25 <- summary(estMI)

return(tabMI)

simulate <- function(nsim, data, seed=41872)

set.seed(seed)

resMCAR25 <- array(NA, dim=c(1, nsim, 10))

for(i in 1:nsim){

data <- data[!is.na(data)]

resMI[1,i,] <- test.impute(data)
}

return(resMI)

resMI <- simulate(nsim)

apply(resMI, c(1,3), mean, na.rm=TRUE)

true <- 0.7024736

biasMI <- mean(resMI[,1] - true)

biasMI

isinMI <- resMI[,6] & true & true | resMI[,7]

covMI <- mean(isinMI)

covMI

int.widthMI <- resMI[,7] - resMI[,6]

average.iwMI <- mean(int.widthMI)

average.iwMI
Appendix B: Imputation Diagnostics

Figure 2: True distribution of the DRG weights.
Figure 3: Distribution of the DRG weights in MI, SI, mean imputation and CCA when the data is MCAR with 25% main diagnosis.
Figure 4: Distribution of the DRG weights in MI, SI, mean imputation and CCA when the data is MCAR with 50% main diagnosis.
Figure 5: Distribution of the DRG weights in MI, SI, mean imputation and CCA when the data is MNAR with 25% main diagnosis.
Figure 6: Distribution of the DRG weights in MI, SI, mean imputation and CCA when the data is MNAR with 50% main diagnosis.
Figure 7: Convergence of DRG weights, MCAR 25%.

Figure 8: Convergence of DRG weights, MCAR 50%.

Figure 9: Convergence of DRG weights, MNAR 25%.

Figure 10: Convergence of DRG weights, MNAR 50%.