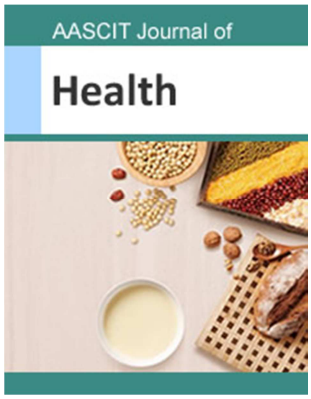
AASCIT Journal of Health
2017; 4(5): 49-57
http://www.aascit.org/journal/health
ISSN: 2381-1277 (Print); ISSN: 2381-1285 (Online)

American Association for Science and Technology



Keywords
Bias, Precision,
Missing at Random,
Unit Level Missing,
Item Level Missing

Received: September 4, 2017
Accepted: November 16, 2017
Published: December 6, 2017

Possibilities of Dealing with Missing Data: An Overview

Jochen Hardt

Medical Psychology and Medical Sociology, Clinic and Policlinic for Psychosomatic Medicine and Psychotherapy, University Medicine Mainz, Mainz, Germany

Email address
hardt@uni-mainz.de

Citation
Jochen Hardt. Possibilities of Dealing with Missing Data: An Overview. *AASCIT Journal of Health*. Vol. 4, No. 5, 2017, pp. 49-57.

Abstract
The way how authors deal with missing data in health care research is often still not optimal, even if modern computers have a high power and there are programs available that would do much better. In the present paper, various ways how to deal with missing data are described, and their pro's and con's are mentioned. Out of the no imputation or single imputation methods, complete case analysis (CC), pairwise deletion (PD), mean imputation, regression imputation, Full information Maximum Likelihood (FIML) and Restricted Maximum Likelihood (REML), hot-deck, missing value indicator, Last observation carried forward (LOCF), Yates method, propensity score and worst case imputation are described. Out of the multiple imputation methods, hot-deck, propensity score, expectation maximation (EM), data augmentation (DA), multiple imputations by chained equations (MICE) and predictive mean matching (PMM) are described. Finally some recommendations were given which method can be applied for which data.

1. Why We do Need to Be Mindful of Missing Data

Missing values in datasets are still widely ignored in health care research. This is unfortunate, because it leads to a loss of precision and sometime even bias in the results. And it is not necessary, because modern computer have enough power to easily handle complex methods for missing values, and a variety of programmes is available. Most statistical methods developed between 1870 and 1970 required data without missing values, so that it became common practice to use only cases where no variables were missing. This led to various drawbacks, which were often overlooked in research practice and are in consequence often not taken into account sufficiently even today.

1.1. Precision

When the method of analyzing complete cases in multivariate procedures is used, even small ratios of missing values can reduce the size of a survey and hence the statistical power of the evaluation. Despite this fact, this is still a default setting in most statistical packages. To illustrate the problem, we will consider a dataset containing a completely random proportion of missing values, about 5% of the cases for each variable. If only complete cases are analyzed, the sample size is reduced by about half if there are 14 variables ($1 - .95^{14} = .52$) if data are missing completely at random (CMAR), i.e. the probability that a case has a missing value in one variable is completely independed of missing values in the other 13 variables. Accordingly, the precision of statistical parameters is often lower than it should be because the available data is only partially used.

1.2. Bias

It has also been shown that restricting the analysis to cases with complete data can lead to a distortion of statistical parameters in certain datasets – so-called bias [e.g. 1]. In most cases, the strength of statistical associations tends to be underestimated, but the opposite can also occur.

A first necessity when dealing with missing data is a distinction between unit non-response and item non-response. Unit non-response is when a subject in a survey does not respond at all, while item non-response means that a subject does not answer one or more questions in a larger set. If only aggregated information about distributions is available, the adequate way to deal with the former is weighting, an overview of this was provided by Seaman et al. [2]. The problem is finding the variables that define the weights. In surveys, age, sex, nationality etc. are generally used. Utilizing them for weighting somewhat improves the estimates, but it is unlikely to remove bias completely because usually additional, unobserved variables contribute to the mechanism of missingness [3]. If register data or background variables are available on an individual level, propensity score adjustment and imputation are alternative options. Here, multiple imputation is one suitable way to handle the missing data. In some cases, however, basic data such as age, sex, etc. are provided from non-responding units, so that item non-response models could be applied to unit non-response. On the other hand, combining both approaches is also a valid option [e.g. 4], this merely entails more complex analyses. An important aspect is that data should be principally observable. For example, it would not make sense to impute the number of pregnancies in men. Extrapolating for cases where all data is missing is also not helpful – some algorithms would leave such cases empty, some would impute a distribution around the sample mean.

For item non-response, a second distinction refers to the mechanisms explaining how missing values occur in a dataset. Data can be Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) [5, 6]. Other classifications [e.g. 7] or further differentiations [e.g. 8] are rarely used nowadays.

MCAR (sometimes CMAR: Completely Missing at Random) means that the pattern of missing data is completely coincidental, i.e. independent of all other variables that can be found in or outside of the dataset. A classic example would be a researcher carrying his questionnaire forms home and being hit by a gust of wind, causing the pages to fly out of his hand and making it impossible for him to retrieve all of them. A missing data pattern can look very regular for example when it is created by design, but the mechanism might still be MCAR.

MAR is a somewhat misleading term because it allows strong dependencies of the mechanism of missing values on other variables in the dataset. If, for example, all the data for men is missing but all the data for women is available, the dataset is MAR as long as sex is contained as a variable. The formal definition states that in order to explain whether a value is missing or observed, the information obtainable from the dataset is sufficient. This definition has the disadvantage that it can never be tested on a real dataset since it is always possible that variables not included in the dataset influence the pattern of missing values, at least partially. Hence, no definitive test exists to prove that data is MAR and approximations can only be performed in large samples [e.g. 9].

MNAR (sometimes also NMAR: Not Missing at Random) means that there is an unknown process in the data accounting for the pattern of missingness. For example, if somebody is interviewing subjects about socially undesirable behaviour like lying, stealing or cheating, it is plausible that missing values imply a higher degree of such behaviour rather than a lower one. Likewise, when assessing quality of life for people with severe, potentially fatal illnesses, there will always be missing values, which, systematically, should be in the lower realm of quality of life. An exact reconstruction of why certain data are missing is generally not possible using only the information provided in the dataset.

2. No Imputation and Single Imputation

Basic methods for substituting missing data were already in use at the beginning of the 20th century. For continuous variables, the mean value was used, for categorical variables the modal value. A more complex method was developed early on by Yates [10], who proposed an iterative method for analysis of variance using line and column means – even before there were computers. With higher computation power in the 1970s, different algorithms to substitute missing data were developed [1]. Two milestones occurred with a book by Rubin alone and a further book by Rubin and Little, both were first published in 1987, the latter has a second edition, now [5, 11]. Little and Rubin (2002) compare various methods of dealing with missing data and demonstrate that multiple imputation performs best in many situations. This chapter describes the mostly applied methods for no imputation and single imputation, even if some of them can also become applied in multiple imputation.

2.1. Listwise Deletion

Listwise Deletion (or CC analysis, the terms refer to identical procedures) is one of the simpler methods of dealing with missing values. People sometimes think ignoring the missing data would not raise a problem. This is wrong, CC only leads to unbiased estimates under MCAR conditions. Also, a reduction of the sample size decreases power and precision. A further popular method was a calculation of covariance, correlations, etc. for those partial datasets for which both variables of a pair were available. This method is referred to as pairwise deletion (PD). Its advantage is that it is more economical in comparison to CC

because more datasets are used without having to perform a substitution. However, the drawback of this method is that individual parameters are estimated on the basis of different partial datasets. In some cases, this may lead to inconsistencies. In an example with only five cases, Arbuckle [12] illustrated that pairwise deletion can lead to a correlation coefficient of r = -1.48, but correlation coefficients larger than |1| are not defined. A further disadvantage of this method becomes apparent when it comes to calculating the standard errors: it is unclear which number of cases should be applied. Pairwise deletion is often used for calculating Cronbach's Alpha (a measure for internal consistency over various items). Here, it's application is relatively unproblematic, because usually items have moderate positive correlations, ideally of similar magnitude.

Simple methods for a substitution, usually referred to as the *ad hoc methods*, replace each missing value with a single, estimated value. The label "ad hoc" was given because they do perform well under certain conditions (usually those under which they were first explored), but not under others. When only few data are missing (say less than 5%), application of these methods usually does not lead to serious distortion. More sophisticated methods have also been developed for single imputation. For example, van Ginkel et al. [13] explored the performance of two methods for substituting missing data when single items from a score had missing values. Both rely on a combination of simple and conditional mean substitution as described below. When up to 15% of the data was missing, almost no bias was seen in the parameters estimated. However, the application of these ad hoc methods is currently out of fashion. Sometimes, reviewers reject articles using ad hoc methods, even though they are appropriate in a certain context. Hopefully, this will change in the future. We will give a brief description of commonly used methods and some of their pros and cons.

2.2. Mean Substitution

The sample mean is substituted for missing values. Formerly, the modal value was used for categorical values, but this has been proven to introduce unforeseeable bias [14] – the mean should therefore also be used to replace missing values in these cases [15]. This sounds easy and often is, but it can also lead to some difficulties. For example, if a variable "sex" is coded "0" and "1", the substituted value may be 0.45. The disadvantage is that a previously binary variable now has three categories, therefore simple statistics like logistic regression can no longer be applied: one would, for example, need to switch to ordered or multinominal logistic regression. Although mean substitution provides undistorted point estimates (means, regression coefficients) in many cases, it has been shown that, in general, the variances are underestimated. Regression coefficients are typically estimated well when missing values in the explanatory variables are substituted by the mean, but they are distorted when data are missing in the responses [e.g. 16, 17] or in both the responses and explanatory variables [18]. A different example would be when substantial amounts of cases in many explanatory variables in small samples receives mean substitution. An analysis can become completely destroyed then.

2.3. Conditional and Stochastic Regression Substitution

Missing values are replaced with the mean of a set of values or a value predicted by a regression equation. Conditional mean substitution provides undistorted point estimates (mean, regression coefficients) when missing data are in the responses, but usually tends to overestimate associations when missing data are in the explanatory variables. Also, it has been shown that variances are also underestimated. As a consequence, if a mediator with missing data is to be analyzed, neither mean substitution nor conditional mean substitution will work, because a mediator is defined as a response for one set of variables, and as an explanatory variable for another set. Correlations are usually biased as well. Hence, the development of more sophisticated methods was desirable. An intuitively appealing approach would be to perform a regression imputation and add some error created by a stochastic process to the estimated values. Enders [19, p. 48] warns against this because there will still be an underestimation of error terms, and with it an increased type I error rate, i.e. falsely significant results. One way to deal with the problem adequately is to generate error terms via bootstrap [19, p. 145ff]. Using graphics, Baraldi and Enders [20] demonstrated vividly what data look like that are generated with mean -, regression - and stochastic regression imputation.

2.4. Expectation Maximization and Full Information Maximum Likelihood

Another method for single substitution of missing values using random error is expectation maximization (EM). With Maximum-Likelihood functions, a covariance matrix is calculated over the incomplete data. Based on this, the value that fits best for any missing item is found. Some random error is added, and then it replaces the missing data. Then, the covariance matrix is fitted again using the imputed values, and the imputed values are again estimated using the new covariance matrix. Again, random error is introduced. This is repeated until a convergence criterion is reached. [21, p. 176] demonstrates on a simple data set how this mechanism converges. Intuitively, the method seems perfect; however, it also underestimates error terms. Enders [19, p. 113] notes that a single EM substitution does not differ much from the stochastic regression approach, the only difference being that the regression prediction is replaced by a maximum likelihood estimate. EM is not an imputation method as such. It works perfectly without estimating the individual values (see next paragraph). The estimated values for individuals are rather a byproduct of the algorithms.

Full information Maximum Likelihood (FIML) and Restricted Maximum Likelihood (REML) are two methods that become increasingly applied. Utilizing a maximum

likelihood approach similar to EM, they can make use of cases having partitially observed data without performing imputations. They have been explored in detail as alternative to Analysis of Variance (ANOVA), in particular with repeated measures [via a procedure called mixed models: 22]. Graham et al. [23] suggest some designs with planned empty cells – impossible to analyze with ANOVA, but with mixed models easy to handle. (1) The data does not need to be balanced, even empty cells can be handled. The former is often referred to as the ability of mixed models to deal with correlated predictors. This is easy to understand. When all cells have the same size, i.e. same number of subjects, the factors are uncorrelated. In experimental designs, this can be striven for and, as long as there are no dropouts, also often achieved. With dropouts or in observational studies, equal cell sizes are often not possible. In these cases, the variance that can be attributed to one factor in ANOVA is not independent of the variance of another factor. There are several ways to deal with it. However, they do lead to different results, which some researchers are probably not aware of (this refers to the so called Type I, II, and III error in Anova). The latter, empty cells, constitute a more or less unsolvable problem in ANOVA. (2) Mixed models allow a specification of the link function between predictors and responses, it can be linear for continuous -, poisson for count - or logit for binary variables. Various other links are available. The link in ANOVA is always linear. Transformation of the responses can partially overcome this problem, but mixed models are more flexible. (3) Mixed models are more efficient than ANOVA when the cells have unequal sizes [24]. (4) Factors comprising more than two categories require contrast or post hoc tests in ANOVA – mixed models set contrast automatically and report coefficients. Contrary to ANOVA, the omnibus tests needs to be requested separately. Mixed models have the advantage of handling data much more easy than multiple imputation, they have the disadvantage that the inclusion of auxiliary variables is more difficult.

It is known that when cells have equal sizes, there are no missing data and data are linked by a linear function, ANOVA and mixed models (using effect coding and the restricted maximum likelihood estimator – REML rather than FIML) return identical results [e.g. 25]. Understanding the mathematics behind mixed models is not easy and the literature describing mixed models is often quite technical. Short descriptions containing only a few formulae are available by Cnaan et al. [26] and Kwok et al. [27]. For a more detailed overview, I would recommend West et al. [28]. Gelman & Hill [29] provide a detailed introduction which is especially interesting for "R" users, Raabe-Hesketh & Skrondal [24] provide a similar one for STATA users, while SPSS users may wish to read Seltman [30].

2.5. Hot Deck

A further group of substitution procedures is called hot deck. The name hot deck derives from the time of punch cards. If a punch card was missing, another was randomly selected from the pile of already read punch cards (hot deck, i.e. still warm from the reading device) and inserted into the reading device. Hence, cases with missing data are substituted with randomly chosen other cases. Hot deck imputations substitute each missing value with the value of an observed neighbour, the latter varies depending on how this is defined. Simple hot deck procedures randomly select a case, more complex processes stratify data based on categorical variables [31, 32]. There are more complex methods which also take continuous variables into consideration [e.g. 33]. As mentioned in the introduction, predictive mean matching can be regarded as a sophisticated hot deck. Probably due to its history, some hot deck procedures not only substitute the missing values, but replace the whole record with one that has no missing values (e.g. STATA's "hotdeck").

2.6. Missing Value Indicator

An intuitively appealing method to deal with missing values is to create a so-called indicator for each variable in a dataset with missing values. This indicator is zero when a value is observed and one when the value is missing. Then, the missing values in the original variables are substituted with a constant. Theoretically, this can be any value, but in most cases it would be the mean. This doubles the number of variables (that had at least one missing observation) in the dataset because for each variable, we have the indicator additionally, but there are no missing values left. Any analysis will now include not only variables of interest but also their missing indicator(s). The latter will always be on the side of the explanatory variables to help decide if an effect is due to the explanatory variable itself or the mechanism of missingness. As easy as this method may seem, it has a serious drawback: coefficients can be distorted. This has been demonstrated in a variety of examples [e.g. 16, 34] and can be explained in several ways. One is to say that the analysis model has changed. If, for example, a simple regression of Y on X was intended and missing values were only in X, we now perform a regression of Y on X and I, with I being the indicator variable of X being missing. The fact that the estimated coefficients b_{YX} and $b_{YX|I}$ will not be identical in both regressions is obvious, but this method still has its place in randomized studies [35]. The missing data indicator may well be utilized to examine the missing data mechanism and is an underlying part of the EM algorithm, but including it simply into the regression equations can lead to distorted coefficients even when the data are fully MCAR.

2.7. Last Observation Carried Forward (LOCF)

A method widely applied in medical longitudinal research was to carry the last observed value forward when a subject dropped out of the study rather than having a missing value. If, for example, a patient received a value of 32 for depression in the second wave of a study and did not respond in the third and following waves, (s)he would keep this value

for all following waves – in many studies without even being contacted again. For continuous variables, the typical statistical approach was to perform a repeated measurement analysis of variance (ANOVA), without considering whether a value was observed or carried forward. Thus, the user obtains a matrix without any missing values and the method even seems to control for different levels among the subjects. Unfortunately, though this method is intuitively appealing, it has two serious disadvantages. First, when group means are changing, the values carried forward do not. Hence, effects can be over- or underestimated by LOCF. If, for example, various cancer treatments for a disease that causes deterioration are to be compared, the one with the most missing data would receive the best evaluation if LOCF is applied. Second, the error term "within subjects" of the ANOVA is reduced, often drastically, if many data are missing. Because treatment effects are usually tested against this error term, LOCF can produce dubious significant results [36, e.g. 37]. The first drawback can be addressed by utilizing a so-called z-LOCF, where a relative position within the group and not the raw value is carried forward [38], but it is unknown today how far this is sufficient in regard to the second. Applying LOCF to binary data is even more problematic [e.g. 39] and cannot be recommended given today's alternative options.[1]

2.8. Yates Method

Early on, Yates [10] suggested a procedure which alternated cycling over rows and columns to estimate values for the missing data in analysis of variance. Though this is also intuitively appealing and works well in the first few cycles, it can run into extreme values when cycling on. It has been described in detail by Little and Rubin [5].

2.9. Propensity Score

For each variable that has missing values, a logistic regression is performed to explain if a value is observed or missing. Explanatory variables are chosen by the analyst. Now, the probability of a variable being missing can be calculated for each case using the regression function, this is called the propensity score. Its distribution is cut into a distinct number of categories, usually five or eight. The missing data in each category is finally replaced with a random draw from the observed data in the same category. This method was developed by Rosenbaum and Rubin [40] to correct for bias in estimating treatment effects. It focuses on missing data in responses and is not recommended for small or medium sample size research [41] or when missing data are in an explanatory variable [42].

2.10. Worst Case Imputation

A method that is widely used in research on alcoholism therapy is the worst case imputation [e.g. 43]. Every missing value is substituted with the worst case, i.e. therapy was not successful or a score of zero. The underlying rationale is that if a subject does not respond after therapy, he has probably relapsed. Apart from the fact that this rationale is quite reasonable sometimes, the method has the added advantage of motivating therapists to collect data – certainly one of the best ways to deal with missing values. On the other hand, it obviously introduces bias [44], and its application is of course restricted to specific cases [45].

3. Multiple Imputation

Multiple imputation creates multiple datasets that contain identical copies of the originally observed data. The missing observations in these datasets are then imputed, using one of the stochastic algorithms described below. Their common ground is that they estimate values utilizing information from the observed values and add random error. Hence, they create different values to be substituted for the missing data in each dataset. The additional variance between various datasets caused by differences in the imputed values reflects the uncertainty of the imputation [46]. The relative variance increase (RVI) due to missing value substitution can be calculated easily for any given multiply imputed dataset [11]. After imputation, statistics are performed separately in all of these datasets and coefficients are combined at the conclusion of the analysis by applying Rubin's rules [11, 47].

Table 1. *Representation of a dataset before and after multiple imputation.*

Original dataset					-	-	**Imputed dataset**					
i	**X_1**	**X_2**	**X_3**	**X_4**	-	-	**i**	**m**	**X_1**	**X_2**	**X_3**	**X_4**
1	0	0	1	2	-	-	1	0	0	0	1	2
-	-	-	-	-	-	-	1	1	0	0	1	2
-	-	-	-	-	-	-	1	2	0	0	1	2
-	-	-	-	-	-	-	1	3	0	0	1	2
2	1	2	mis	4	-	-	2	0	1	2	mis	4
-	-	-	-	-	-	-	2	1	1	2	2	4
-	-	-	-	-	-	-	2	2	1	2	1	4
-	-	-	-	-	-	-	2	3	1	2	3	4
3	2	mis	mis	3	-	-	3	0	2	mis	mis	3
-	-	-	-	-	-	-	3	1	2	2	2	3
-	-	-	-	-	-	-	3	2	2	1	1	3
-	-	-	-	-	-	-	3	3	2	2	3	3
...	-	-	-	-	-	-	3	-	-	-	-	-
N	mis	1	3	mis	-	-	N	0	mis	1	3	mis
-	-	-	-	-	-	-	N	1	3	1	3	3
-	-	-	-	-	-	-	N	2	4	1	3	2
-	-	-	-	-	-	-	N	3	3	1	3	2

To provide an example: with three imputed datasets, a result will generally appear as displayed in Table 1. In this case, "i" is the individual identifier, "m" the imputation number, X_1 to X_4 are the variables and "mis" indicates missing values. Most programs will copy the original dataset with the number m = 0 into the imputed dataset. This is practical because it allows the user to review and analyze the

1 A variant of LOCF is Baseline Observation Carried Forward (BOCF). It is sometimes applied with the rationale that it is very conservative, so cannot do harm. Such a rationale is wrong, any critique on LOCF described above also holds true for BOCF, it may even perform worse than LOCF.

original data in the program without much effort. Lines with m = 0 are not incorporated into the MI analyses[2]. It can be seen that in the first case, which had no missing data, all data were adopted identically in the three imputed datasets. In the second case, there are substitutions for X_3, which differ in the three datasets. The values of the other three variables are again adopted identically [e.g. 48]. Various algorithms exist to determine the imputed values.

3.1. Hot Deck

Early on, Hot Deck methods were used for single imputations. Multiple imputations are obtained if the operation is repeated several times [even using programs that are not designated for multiple imputations, e.g. 31, 49]. In first simulations with medium-size datasets (n=500), this led to good results [e.g. 50] and hot deck is still recommended by some researchers [51]. Others compared it with various other methods of multiple imputation and found that it was usually not the best-suited method [e.g. 14, 52]. In some own simulations, hot deck did not perform well when the proportion of missing data was large. Since other procedures which require the same effort but perform better are available, I will not discuss hot deck further.

3.2. Propensity Score Estimation

Propensity Score Estimation can also be repeated to perform multiple imputation. Unlike predictive mean matching, which directly estimates a value and replaces with the nearest neighbour, in propensity score matching it is not the missing value itself that is estimated first, but the fact that a value is missing or not. This probability creates the donor pool where the value to replace a missing data is drawn from. It was examined in detail by Allison [42] and did not lead to a satisfactory result. Allison states that it produces "badly biased estimates when data on a predictor variable are missing at random or even missing completely at random" (p. 301). Solas [53] offers two multiple imputation routines utilizing the propensity score estimate.

3.3. Expectation Maximization

Expectation Maximization (EM) algorithms [54], were initially utilized for single imputations, but multiple imputation leads to better results here as well. If the process is performed repeatedly with different random values, a variation in the substitutions occurs. Implementations of EM algorithms differ significantly in individual programs. Amelia II, for example, applies a Bootstrapping-Sampling in addition to EM [55]. It should be noted that these algorithms were developed for continuous variables, handling categorical ones requires some additional care

3.4. Data Augmentation (DA)

Data Augmentation (DA) works similar to EM, but relies on stochastic processes instead of the (deterministic) maximum likelihood function using a special version of Gibbs sampling [21, pp 181-9]. EM and DA share the fact that parameters for the variables of a data set are assessed jointly, therefore they are both called Joined Modelling (JM). It makes them extremely fast running on the computers. Additionally they have the advantage of a better theoretical foundation [56, p. 116] then fully conditional specification (see paragraph below), but the main disadvantage is that they were developed for quantitative variables.

3.5. Multiple Imputations by Chained Equations

Multiple Imputations by Chained Equations [MICE: 1] is an algorithm used increasingly often. It is also referred to as fully conditional specification (FCS). First, based on a random sample from the data, a regression equation is calculated in order to predict missing values in any variable. Depending on the scale level of the variable containing missing data that is to be substituted, it can basically be a linear, logistic, ordered logistic or multi-nominal regression. Random error is also added to the estimation. The program then switches to the next variable, and imputes the missing data there. Contrary to EM, the missing values in the variables are not imputed simultaneously, but step by step while cycling over the variables. Similar to EM, MICE algorithms circulate over all variables several times. During this process, the estimated values of the previous substitutions are the starting values for the next series of substitutions. For MICE, defining a convergence criterion is difficult, therefore a definite number of cycles (usually 10) is specified. More detailed descriptions of the MICE algorithm are provided by Allison [57] or van Buuren [56, 58]. Differences regarding the performance of various MICE algorithms are smaller than those between MICE and EM algorithms, but they are indeed present. For example, the sequence of the variables that are to receive imputed values differs, and most importantly how they deal with a situation when a categorical variable receives a perfect prediction.

3.6. Chained Equation Algorithms

Chained equation algorithms can be combined with predictive mean matching (PMM). Here, the regression equation to predict the missing values for a certain variable is taken. Cases with observed and missing values are sorted according to their predicted value (ranking). Any case without an observed value will adopt the observed value from its nearest neighbour on the prediction line[3]. If there is

2 In fact, there are some differences between and within the programs regarding how multiply imputed data can be stored. The above described structure is an example for one that is easy to understand. Especially when transferring imputed data from one program to another, it becomes necessary to check the storing format.

3 In fact, the selection isn't limited to one single neighbor. Rather, a group of the size "k" (usually 3-5) should be chosen. Of these, one is selected randomly to "donate" its observed value. The procedure is also called "k nearest neighbors" and produces more stable estimates than selecting just one neighbor.

more than one nearest neighbour, a random choice is drawn. Using PMM after chained equations is said to be robust against violations of assumptions of MI, such as multivariate normality [59-61]. It solves some problems other algorithms face, such as rounding and setting boundaries. The chained equations algorithm combined with PMM is probably the least problematic way to perform multiple imputations – on the cost of long computing times.

After multiple imputation, the statistical parameters are estimated separately for the various imputation samples and combined. For this combination, Rubin's rules are usually applied [11]. Even if the programs are getting better, imputation itself and analysis of the multiply imputed data set can cost considerably more time than analyzing a single data set. Theoretically, any statistic can be done in multiply imputed data sets, practically programs offers are limited. But in addition to the usual statistical parameters, in multiply imputed data, the relative variance increase (RVI), which is the ratio of variances between and within the m datasets, is computed for a certain estimate. It is small when there are few missing data and has a maximum of 1 when all variation is due to the imputation – which will rarely be the case. A related measure, which is sometimes confused with the RVI, is the fraction of missing information [FMI: 11, 56, p 41]. When there is only one variable with missing data and there is an MCAR mechanism, it would be identical to the proportion of missing data. When more than one variable has missing data and these variables are correlated, it usually becomes smaller because some missing information can be estimated based on neighboring variables.

Degrees of freedom (DF) for the denominator are adjusted using a ratio of variances between and within the imputations [e.g. 19]. When this ratio is small, the denominater-DF can be larger than the sample size, though this may seem counterintuitive. The formulae were suggested by Rubin and Schenker [62] and proven to be valid in large samples, while some modifications were later suggested for smaller samples [63, 64]. Van Buuren [56, pp 42f] briefly describes Barnard and Rubin's [63] formula to adjust the degrees of freedom so that their number is never higher than the number of cases. Today, most programs may still provide DFs that are larger than the sample size. Unless the sample is not extremely small, it should not bother a researcher too much – it is mainly a problem of aesthetics.

It is agreed upon that DA and MICE lead to very similar results in large samples, especially when many datasets are created for multiple imputations [e.g. 65, 66]. For multiple hot deck imputations, there are not enough examples to allow for a comparison to DA or MICE.

4. Conclusion and Recommendations

I described some of the mostly applied methods to deal with missing data, to allow the reader to make a choice which one to apply in a certain situation. Here, I want to conclude with four warnings. (1) Any use of LOCF or BOCF today should be accompanied by a clear rationale why it is superior to less problematic methods in the specific case. (2) Applying the indicator variable method should be very well considered, taking into account that beside dealing with the missings it comprises a change of the measurement model. (3) With hot-deck procedures it should be kept in mind that not only the missing data are substituted, but also some of the observed ones. (3) And the propensity score method is surely for experienced statisticians only, a wrong implementation can do more harm than good.

However, beside personal preferences, characteristics of the data set count. If there are very few missings in a data set (< 1%), there are not many variables and the mechanism is likely to be mainly MCAR, CC is usually fully sufficient. If there are more missings (< 5%), a simple substitution like the mean will often be a good solution – it enhances the power with a risk of (a) introducing small bias and (b) making significance tests falsely positive. Both effects are, given the small amount of missing data, in most cases negligible. Handbooks of psychological tests often recommend to substitute single missing items of a scale by the mean of the other items. If items have similar means and less than $^1/_3$ of the items of the respective scale are missing, this seems to be a good advice. If item means vary strongly, the sample mean would be a better choice. If the proportion of missings is larger than 5%, van Ginkels [13] method can be applied to item sets, and mixed models can substitute ANOVA or regression models for other variables. Van Ginkels method makes use of the information that non-missing variables of the data set measuring the same construct carry, but mixed models do not. If such information is assumed to be present in the data set, multiple imputation would become an option and the additional effort of performing separate analyses and combining them is likely to pay off. When all variables are continuous and approximately normally distributed, I would chose DA due to computational speed. When data are (partly) categorical or strongly skew, MICE in combination with PMM would be my first choice.

References

[1] Rubin, D. B., *Inference and missing data.* Biometrica, 1976. 63: p. 581-592.

[2] Seaman, S. R. and White, I. R., *Review of inverse probability weighting for dealing with missing data.* Stat Methods Med Res, 2011.

[3] Schneider, K. L., Clark, M. A., Rakowski, W., and Lapane, K. L., *Evaluating the impact of non-response bias in the Behavioral Risk Factor Surveillance System (BRFSS).* J Epidemiol Community Health, 2012. 66 (4): p. 290-5.

[4] Seaman, S. R., White, I. R., Copas, A. J., and Li, L., *Combining multiple imputation and inverse-probability weighting.* Biometrics, 2012. 68: p. 129-37.

[5] Little, R. J. and Rubin, D. B., *Statistical Analysis with Missing Data*. 2002, New York: Wiley.

[6] Rubin, D. B., *Multiple imputations after 18 plus years.* JASA, 1996. 91: p. 473-89.

[7] Winship, C. and Mare, R. D., *Models for sample selection bias.* Annual Review of Sociology, 1992. 18: p. 327-350.

[8] Potthoff, R. F., Tudor, G. E., Pieper, K. S., and Hasselblad, V., *Can one assess wether missing data are missing at random in medical research.* Stat Methods Med Res, 2006. 15: p. 213-234.

[9] Jamshidian, M. and Jalal, S., *Tests of homoscedasticity, normality and missing at random for incomplete multivariate data.* Psychometrika, 2010. 75 (4): p. 649-74.

[10] Yates, F., *The analysis of replicated experiments when the field results are incomplete.* Emporium Journal of Experimental Agriculture, 1933. 1: p. 129-42.

[11] Rubin, D. B., *Multiple imputation for nonresponse in surveys*. 1987, New York: Wiley & Sons.

[12] Arbuckle, J., *Full information estimation in the presence of missing data*, in *Advanced Structural Modelling*, Marcoulides, G. A. and Schumaker, R. E., Editors. 2004, Erlbaum: NJ.

[13] van Ginkel, J. R., van der Ark, L. A., and Sijtsma, K., *Multiple imputation of item scores in test and questionaire data, and their influence on psychomertic results.* Multivariate Behavioural Research, 2007. 42 (2): p. 387-414.

[14] Ambler, G., Omar, R. Z., and Royston, P., *A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome.* Statistical Methods in Medical Research, 2007. 16 (3): p. 277-98.

[15] Schemper, M. and Heinze, G., *Probability imputation revisited for prognostic factor studies.* Stat Med, 1997. 16 (1-3): p. 73-80.

[16] Donders, A. R., van der Heijden, G. J., Stijnen, T., and Moons, K. G., *Review: a gentle introduction to imputation of missing values.* J Clin Epidemiol, 2006. 59 (10): p. 1087-91.

[17] Schafer, J. L. and Graham, J. W., *Missing data: our view of the state of the art.* Psychological Methods, 2002. 7: p. 147-177.

[18] Schlomer, G. L., Bauman, S., and Card, N. A., *Best practices for missing data management in counseling psychology.* J Couns Psychol, 2010. 57 (1): p. 1-10.

[19] Enders, C. E., *Applied missing data analysis*. 2010, New York: Guilford.

[20] Baraldi, A. N. and Enders, C. K., *An introduction to modern missing data analyses.* J Sch Psychol, 2010. 48 (1): p. 5-37.

[21] Schafer, J. L., *Analysis of incomplete multivariate data*. 1997, New York: CRC Press.

[22] Ferro, M. A., *Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood.* Annals of Epidemiology, 2014. 24: p. 75-77.

[23] Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E., *Planned missing data designs in psychological research.* Psychol Meth, 2006. 111: p. 323-43.

[24] Raabe-Hesketh, S. and Skrondal, A., *Multilevel and longitudinal modelling using Stata*. 2005, College Station, TX: Stata Press.

[25] Laird, N. M. and Ware, J. H., *Random-effects models for longitudinal data.* Biometrics, 1982. 38 (4): p. 963-74.

[26] Cnaan, A., Laird, N. M., and Slasor, P., *Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data.* Stat Med, 1997. 16 (20): p. 2349-80.

[27] Kwok, O. M., Underhill, A. T., Berry, J. W., Luo, W., Elliot, T. R., et al., *Analyzing longitudinal data with multilevel models: An example with individuals living with lower level extremety intra-artcular fractures.* Rehabil Psychol, 2008. 53 (3): p. 370-86.

[28] West, T. B., Welch, K. B., and Galecki, A. T., *Linear mixed models: a practical guide using statistical software*. 2007, Boca Raton: Chapman & Hall.

[29] Gelman, A. and Hill, J., *Data analysis using regression and multilevel/hierarchical models*. 2006, New York: Cambridge University Press.

[30] Seltman, H., *Experimental Design for Behavioral and Social Sciences*, ed. http://www.stat.cmu.edu/~hseltman/309/. 2013.

[31] Little, R. J., Yosef, M., Cain, K. C., Nan, B., and Harlow, S. D., *A hot-deck multiple imputation procedure for gaps in longitudinal data on recurrent events.* Stat Med, 2008. 27 (1): p. 103-20.

[32] Siddique, J. and Belin, T. R., *Multiple imputation using an iterative hot-deck with distance-based donor selection.* Stat Med, 2008. 27 (1): p. 83-102.

[33] Andridge, R. R. and Little, R. J., *A review of Hot Deck imputation for survey response.* International Statistical Review, 2010. 78: p. 40-64.

[34] Knol, M. J., Janssen, K. J., Donders, A. R., Egberts, A. C., Heerdink, E. R., et al., *Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example.* Journal of Clinical Epidemiology, 2010. 63 (7): p. 728-36.

[35] Groenwold, R. H., White, I. R., Donders, A. R., Carpenter, J. R., Altman, D. G., et al., *Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis.* Cmaj, 2012.

[36] Salim, A., Mackinnon, A., Christensen, H., and Griffiths, K., *Comparison of data analysis strategies for intent-to-treat analysis in pre-test-post-test designs with substantial dropout rates.* Psychiatry Res, 2008. 160 (3): p. 335-45.

[37] Olsen, M. K., Stechuchak, K. M., Edinger, J. D., Ulmer, C. S., and Woolson, R. F., *Move over LOCF: principled methods for handling missing data in sleep disorder trials.* Sleep Med, 2011. 13 (2): p. 123-32.

[38] Hendrix, S. B. and Wilcock, G. K., *What we have learned from the myriad trials.* J Nutr Health Aging, 2009. 13 (4): p. 362-4.

[39] Cook, R. J., Zeng, L., and Yi, G. Y., *Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation.* Biometrics, 2004. 60 (3): p. 820-8.

[40] Rosenbaum, P. R. and Rubin, D. B., *The central role of the propensity score in observational studies for causal effects.* Biometrica, 1983. 70 (1): p. 41-55.

[41] Pearl, J., *Remarks on the method of propensity score.* Stat Med, 2009. 28: p. 1415-1423.

[42] Allison, P. D., *Multiple imputation for missing data: a cautionary tale.* Sociological Methods & Research, 2000. 28: p. 301 - 9.

[43] McPherson, S., Barbosa-Leiker, C., Burns, G. L., Howell, D., and Roll, J., *Missing data in substance abuse treatment research: current methods and modern approaches.* Exp Clin Psychopharmacol, 2012. 20 (3): p. 243-50.

[44] Hardouin, J. B., Conroy, R., and Sebille, V., *Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data.* BMC Med Res Methodol, 2011. 11: p. 105.

[45] Hallgren, K. A. and Witkiewitz, K., *Missing Data in Alcohol Clinical Trials: A Comparison of Methods.* Alcoholism-Clinical and Experimental Research, 2013. 37 (12): p. 2152-2160.

[46] Rubin, D. B., *Multiple imputation in sample surveys.* Proc Survey Res Meth Sec Am Statist Assoc, 1978. 20-34.

[47] Marshall, A., Altman, D. G., Holder, R. L., and Royston, P., *Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines.* BMC Medical Research Methodology, 2009. 9: p. 57.

[48] Enders, C. K., *A primer on the use of modern missing-data methods in psychosomatic medicine research.* Psychosom Med, 2006. 68 (3): p. 427-36.

[49] Reilly, M., *Data Ananysis using Hot Deck multiple imputation.* Journal of the Royal Statistical Society. Series D, 1992. 42: p. 307 - 313.

[50] Reilly, M. and Pepe, M., *The relationship between hot-deck multiple imputation and weighted likelihood.* Stat Med, 1997. 16 (1-3): p. 5-19.

[51] Wang, C. N., Little, R., Nan, B., and Harlow, S. D., *A hot-deck multiple imputation procedure for gaps in longitudinal recurrent event histories.* Biometrics, 2011. 67 (4): p. 1573-82.

[52] Horton, N. J. and Kleinman, K. P., *Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models.* Am Stat, 2007. 61 (1): p. 79-90.

[53] Solas, *Statistical Solutions.* http://www.statsols.com/propensity-score-based-multiple-imputation/, 2014.

[54] Dempster, A. P., Laird, N., and Rubin, D. B., *Maximum Likelyhood from incomplete data using the EM algorithm.* J R Stat Soc (Series B), 1977. 39: p. 1-38.

[55] Honaker, J. and King, G., *What to do about missing values in time-series cross sectional data.* American Journal of Political Science, 2010. 54: p. 561-581.

[56] van Buuren, S., *Flexible imputation of missing data*. 2012, Boca Raton: CRC Press (Chapman & Hall).

[57] Allison, P. D., *Multiple Imputation for Missing Data: A Cautionary Tale.* Sociolocical Methods & Research, 2000. 28 (3): p. 301-9.

[58] van Buuren, S., Boshuizen, H. C., and Knook, D. L., *Multiple imputation of missing blood pressure covariates in survival analysis.* Statistics in Medicine, 1999. 18 (6): p. 681-94.

[59] Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., et al., *Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses.* J Clin Epidemiol, 2002. 55 (2): p. 184-91.

[60] Marshall, A., Altman, D. G., and Holder, R. L., *Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study.* BMC Med Res Methodol, 2010. 10: p. 112.

[61] Landerman, L., Land, K. C., and Pieper, C. F., *An empirical evaluation of the predictive mean matching method for imputing missing values.* Sociolocical Methods & Research, 1997. 26: p. 3-33.

[62] Rubin, D. B. and Schenker, N., *Multiple imputation for interval estimation from simple random samples with ignorable nonresponse.* J American Statistical Association, 1986. 81: p. 366-74.

[63] Barnard, J. and Rubin, D. B., *Small-sample degrees of freedom with multiple imputation.* Biometrika, 1999. 86 (4): p. 948-55.

[64] Lipsitz, S. R., Parzen, M., and Zhao, L. P., *A degrees-of-approximation in multiple imputation.* J Statist Comput Simul, 2002. 72 (4): p. 309-18.

[65] Collins, L. M., Schafer, J. L., and Kam, C. - M., *A comparison of inclusive and restrictive strategies in modern missing data procedures.* Psychological Methods, 2001. 6: p. 330-51.

[66] Lee, K. J. and Carlin, J. B., *Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation.* Am J Epidemiol, 2010. 171 (5): p. 624-32.