

## CHOOSING APPROPRIATE METHODS FOR MISSING DATA IN MEDICAL RESEARCH: A DECISION ALGORITHM ON METHODS FOR MISSING DATA

### Hatice UENAL

Institute of Epidemiology and Medical Biometry, Ulm University , Germany



### Benjamin MAYER

Institute of Epidemiology and Medical Biometry, Ulm University , Germany



E-mail: benjamin.mayer@uni-ulm.de

### Jean-Baptist DU PREL

Institute of History, Theory and Ethics of Medicine, Ulm University, Germany



*\*HU and BM contributed equally.*

### Abstract

Missing data (MD) are a common problem in medical research. When ignored or treated not appropriately, MD can lead to seriously biased results. Currently, there are no comprehensive guidelines for efficiently identifying suitable imputation methods in different MD situations. The objective of the paper is to discuss various methods to handle missing data. Based on a selective literature search, common MD imputation methods were identified. A decision algorithm is presented where the considered methods are prioritized with respect to the underlying missing data mechanism and scale level of the incomplete data. Furthermore, all included imputation methods are described in more detail. No alternative decision algorithms for MD imputation methods of this complexity have been developed yet, thus it could serve as a useful tool for researchers confronted with MD.

**Keywords:** Missing data, decision algorithm, imputation methods, multiple imputation

### Introduction

Missing values are a common problem in medical research and may be a considerable source of bias when ignored or not handled appropriately (Schafer & Graham, 2002). Most studies come along with some type of missing data (questionnaire data, laboratory values, missing records, subject loss to follow-up, etc.). There is often no information in publications which measures were taken into account for missing data (MD) and what impact these choices may had on the results (Eekhout et al., 2012). Furthermore, there are few guidelines in epidemiology as well as in clinical research (Little et al., 2012) which explicitly focus on the MD problem.

The capabilities of imputation might be unknown or inappropriately used for minimizing effects (reduction of power, introduction of bias) of MD on study results. A recently published systematic review on studies in three leading epidemiological journals showed that complete case analysis (CCA) (81%) and single imputation methods (15%) were most frequently used to cope with MD (Eekhout et al., 2012). Both approaches are easily applicable but often not appropriate to deal with MD, as shown in the following. In many cases the adverse effect of MD on a study's validity might not be fully aware and those who are aware may not know that validity could be improved with imputation. Finally, those who wish to use imputation may not know which method is suitable in a particular situation.

Currently, there is no detailed guidance which provides researchers from various disciplines analyzing medical data with information on choosing a proper MD technique for their analysis. In order to address this gap, a comprehensive decision-making algorithm for the application of established MD imputation methods is suggested. Statistical imputation is a powerful tool to handle with MD, so researchers have to be encouraged to deal with the available methods in order to get more reliable study results. The presented algorithm is able to facilitate this process.

**Methods**

This article is based on the assumption that missing values are only present in a single variable (in the following called "target variable"), which can be any variable of the dataset including the outcome variable. Classifying the cause of missing values is of fundamental importance for determining how they should be handled. Therefore, the developed decision algorithm included a distinction between established missing data mechanisms (Table 1) which were initially described by Rubin (1976). A Missing Completely At Random (MCAR) mechanism is present if the missing values of a data set are a random sub-sample of the complete data set, i.e. the probability of MD is independent of all other variables in the data set including the target variable. An example of MCAR may be a patient who dies in a traffic accident during the course of a clinical trial. Assuming a Missing At Random (MAR) mechanism, the probability of MD depends on other variables of the data set, but not on the values of the target variable. MAR is present, for example, if gender predicts the probability of response on a depression score. In case of Not Missing At Random (NMAR), the probability of MD depends on the observed as well as unobserved variables. A NMAR is present if a subject with manifest depression will probably not report about his mental constitution since he fears the consequences of doing so (e.g. inpatient treatment).

**Table 1.** Overview of missing data mechanisms

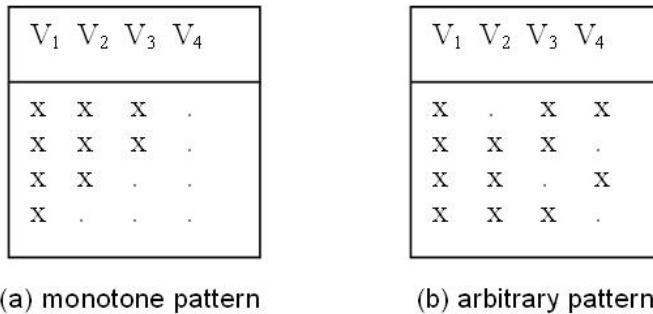
<b>MCAR<sup>a</sup></b>	<b>MAR<sup>b</sup></b>	<b>NMAR<sup>c</sup></b>
Probability of MD is unrelated to covariates and the values of the target variable	Probability of MD could be related to covariates, but not to values of the target variable	Probability of MD relates to unobserved values of the target variable even after control for covariates

<sup>a</sup> Missing Completely At Random <sup>b</sup> Missing At Random <sup>c</sup> Not Missing At Random; MD=missing data

It is usually not easy to distinguish between these three concepts, so a sound knowledge on substantial coherences in the data set is necessary. The mechanisms do not always provide a logical-causal explanation for the absence of data, nevertheless they offer

a mathematical approach to model the probability of MD in association with other variables in the dataset.

In longitudinal studies with MD in a single variable (or studies with MD in multiple variables in general), it is also necessary to look for patterns generated by MD in the data set. Monotone patterns have to be distinguished from arbitrary patterns (Figure 1), since the shape of the pattern affects the applicability of particular imputation approaches (Schafer & Graham, 2002). Hence, the decision algorithm also considers the missing data pattern as a relevant criterion for choosing appropriate imputation methods.



**Figure 1.** Missing data patterns

Along with the missing data mechanism and pattern the third important aspect to be considered within a decision algorithm on MD methods is the scale level of the variable to be imputed. There are methods for categorical as well as continuous variables which have to be chosen properly.

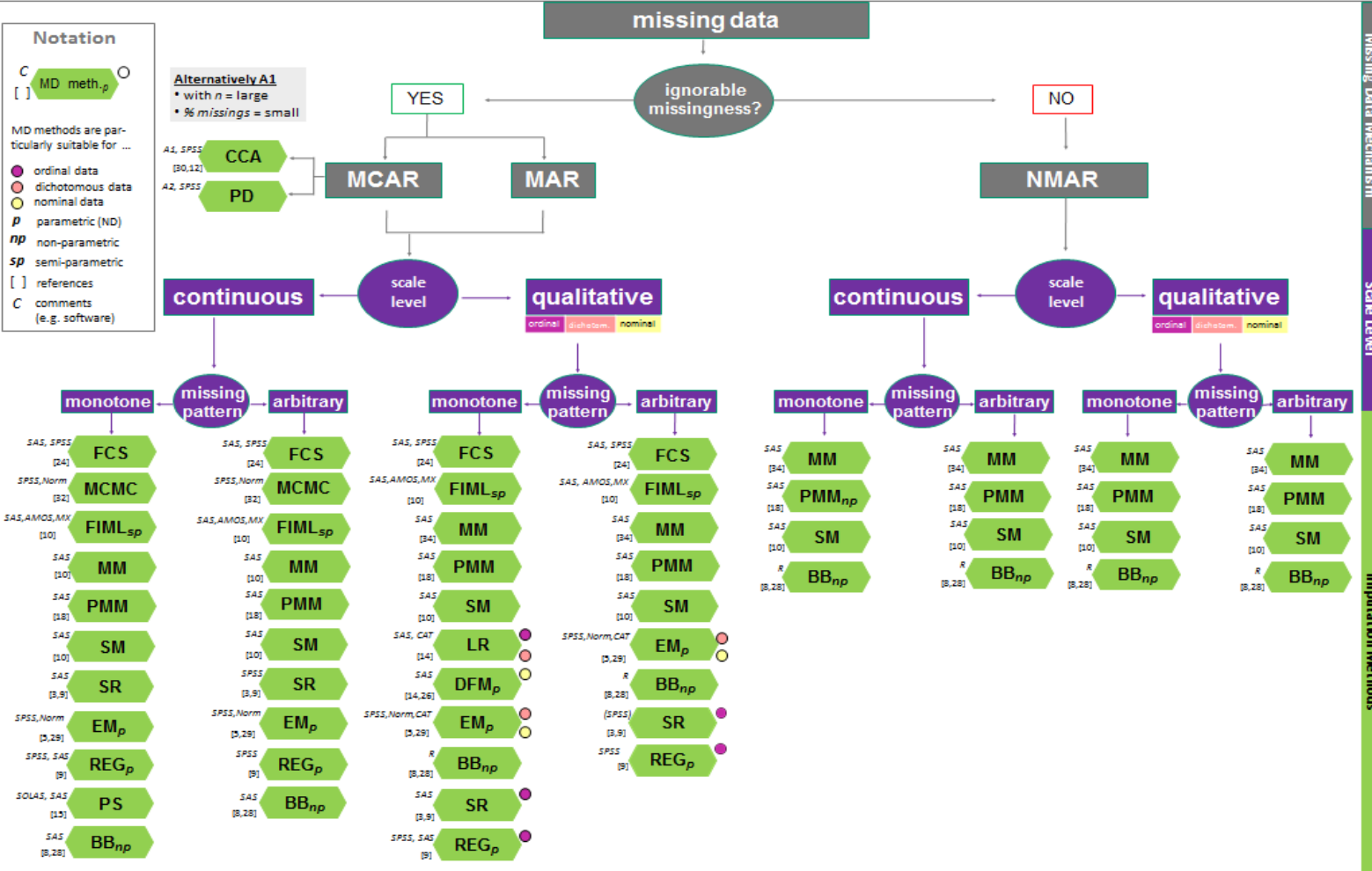
Common parametric as well as non-parametric imputation procedures were identified by selective literature search. Based on the available information in the literature, criteria for choosing adequate imputation methods were discussed and prioritized. They formed the base for the development of the proposed decision algorithm. The prioritization of methods rely on the findings of referenced studies which compared different imputation methods.

**Results**

The above described criteria for choosing appropriate methods to handle with MD are considered within the proposed decision algorithm (Figure 2). This guidance can efficiently assist researchers in finding appropriate imputation methods. To illustrate the use of the proposed algorithm, the following hypothetical MD situations are exemplarily considered: (i) Following a one-time survey of smoking behaviour, the data did not reveal a clear pattern of MD in the variable "current use (few, a packet or more than a packet of cigarettes per day)" and (ii) it was previously known that in a longitudinal psychiatric study patients with diagnosed depression are more likely to have MD due to a reduced willingness to report about their current mental condition (score value out of 5 items).

Considering these scenarios, an initial examination whether the MD have occurred randomly or systematically is required. Given the definitions in Table 1, the mechanisms MCAR and MAR indicate situations where the probability of MD are assumed to be non-systematic. Therefore, at least a MAR mechanism can be assumed for situation (i), since the

probability of MD in the variable "current use" seems to be independent of the covariate structure of the respective patients. In contrast, a NMAR mechanism is most likely for situation (ii) since patients with previously diagnosed depression are more likely to deny reporting about their actual mental condition. Moreover, the target variable in (i) is "current use" and according to the description above an ordinal scaled qualitative outcome. The target variable "mental condition" in (ii) is a score value derived from five items, thus it can be considered as a continuous variable. Further assumptions are that in situation (i) there is an arbitrary MD pattern and in situation (ii) it is possible to induce a monotone MD pattern.



**Figure 2.** Decision algorithm on missing data methods

Applying the presented MD decision algorithm (Figure 2), a full information maximum likelihood (FIML) procedure might be adequate to impute ordinal categorical data in (i) (Enders & Bandalos, 2010). In situation (ii) a mixed model or pattern mixture model to handle the missing score values may be appropriate (Enders & Bandalos, 2010, Verbeke & Molenberghs, 2000).

*Imputation strategies and methods*

In the following, a summary of the methods presented in Figure 2 is provided. An in-depth description of all methods is out of the scope of this article, but interested readers may use the cited references given in the algorithm for each suggested method to obtain more information.

Complete case analysis (CCA) discards subjects with incomplete data in any variable from analysis (listwise deletion) (Enders, 2010). This may be justifiable only in the MCAR situation when complete cases are a random sample of all cases (Vach & Blettner, 1991), but generally problematic in MAR or NMAR situations. The loss of statistical power increases as the number of MD is high. Cases with MD in at least one variable are common, especially in epidemiological studies, where it is typical to have a large number of variables. Hence, even an overall small amount of MD can lead to a dramatic reduction of evaluable cases (Figure 3).

Variable	1	2	3	...	...	...	p	
1	✓	✓	✓	✓	⚡	✓	⚡	✓
⋮	⚡	✓	✓	✓	✓	✓	✓	✓
Subject	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓							
⋮	✓	✓	✓	✓	✓	✓	⚡	✓
n	✓	⚡	✓	✓	✓	✓	✓	✓

**Figure 3.** Loss of information with a complete case analysis

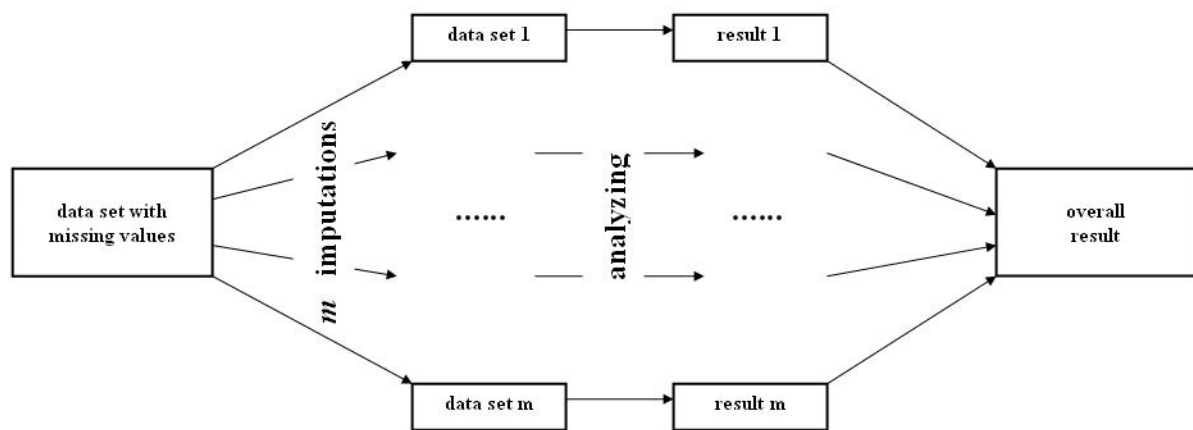
Available case analysis (pairwise deletion) is another method which may be applied in the MCAR situation if the number of MD is low. It overcomes the problem of high loss of cases and power due to MD in contrast to CCA, but has other disadvantages (Enders, 2010). This method utilizes all available data of the subsample used for the actual analysis which contains all observation units with no MD in the variables included. Therefore it is possible that every analysis including different variables has a different number of cases. This difference becomes especially obvious when calculating for example correlation coefficients on the base of subsamples with different numbers in a covariance matrix. Then correlation coefficients of over 1 could result (Enders, 2010).

Single Imputation (SI) methods are characterized by replacing each missing value of a dataset by just one imputation value. There are numerous SI methods and all of them have the tendency to underestimate variance, because the imputation itself is not explicitly considered within the imputation as an additional source of variance. Most of these approaches are not recommended and are therefore not included in the decision algorithm. The most common SI-methods are mean or median imputation and so called "hot and cold" deck imputation methods. Regression based imputation uses information of the complete cases to estimate MD in incomplete variables. The regression approach shares this "bond" of available data for MD estimation with maximum likelihood methods, though the latter use more complicated algorithms (Enders, 2010).

A more sophisticated technique is the stochastic regression imputation which overcomes the disadvantage of the simple regression imputation of predictable bias due to a lack of variability (Baraldi & Enders, 2010). Similar to the simple regression approach, this method imputes the MD of a variable according to their relationship to other covariates, but additionally incorporating a normally distributed residual to account for the reduced variability of the imputed variables. Simulations have shown that stochastic regression provides reasonable results similar to those of Multiple Imputation methods or maximum likelihood methods (Enders, 2010).

Maximum likelihood estimation is another approach to handle MD (Schafer & Graham, 2002). The Expectation Maximisation algorithm (EM) iterates maximum likelihood estimations (repeated, stepwise approximations) of a set of regression equations for predicting MD to get the best solution (Dempster, Laird, & Rubin, 1977). Another modern maximum likelihood based imputation is the full information maximum likelihood method (FIML) which is based on structure equation models and takes all available information into account (Enders & Bandalos, 2001).

The Multiple Imputation (MI) strategy incorporates several state of the art techniques. A MI analysis includes three distinct phases, the imputation phase (1), the analysis phase (2), and the pooling phase (3) (Figure 4). Initially, missing values are imputed  $m$  times ( $m > 1$ ), resulting in  $m$  completed data sets (Rubin, 1987). Each of these datasets is analyzed separately afterwards according to the statistical model chosen to answer the research question. The  $m$  single results (e.g. an estimated regression coefficient) are finally combined to one MI estimator (Little & Rubin, 2002).



**Figure 4.** The principle of Multiple Imputation

From a statistical point of view, MI is able to approximate the observed data likelihood by the average of the completed data likelihood over unknown missing values (He, 2010). MI produces more reasonable estimates than SI procedures (Tanner & Wong, 1987) because the imputation process is considered as an additional source of variation.

The Markov Chain Monte Carlo (MCMC) method uses available information from the observed data to calculate a corresponding a-posteriori distribution with the help of different algorithms, e.g. data augmentation (Tanner & Wong, 1987). Here, likelihood-based sampling for the imputation of MD is performed. It is a quite robust MI procedure for quantitative variables which can be applied to arbitrary MD patterns and produces valid values even for high proportions of MD in the case of MCAR or MAR.

An alternative algorithm for more complex data structures, especially when multiple variables have MD, is known as fully conditional specification (FCS) or multiple imputation by chained equations (MICE). It enables the imputation of quantitative as well as qualitative variables (Raghunathan et al., 2001). Imputation proceeds from one variable with the least to that with the most MD. At each step random draws are made from both the posterior distribution of the parameters and the posterior distribution of the missing values. Imputed values at one step are used as predictors in the imputation equation at subsequent steps. Once all MD have been imputed, several iterations of the process are repeated before selecting a completed data set.

The Propensity Score method is conceptualized for a MI of continuous variables and demands a monotonous MD pattern (Lavori, Dawson, & Shera, 1995). In the context of MD, the propensity score represents a conditional probability that a subject has a missing value given a structure of covariates (Lavori et al., 1995). A propensity score is calculated based on a logistic regression model for each subject and all subjects of a data set are then sorted in an ascending order and distributed to a number of subsets according to their corresponding propensity score. The MD is imputed by a random draw of observed values of a certain subset.

The logistic regression procedure is an appropriate MI approach for binary or ordinal variables with MD in the MCAR or MAR situation (Hohl, 2007; Sentas, Angelis, & Stamelos, 2013; Rubin, 1987). A monotonous MD pattern is required. First, a regression model is set up with all complete cases to estimate the regression coefficients. For given covariates the probability  $p$  is calculated that one of two possible values 0 or 1 is used for imputation of a missing value. Comparing  $p$  with a uniformly distributed random variable  $u$ , a decision is made if the imputation value is 0 or 1.

The discriminant function method imputes MD of a nominal variables and assumes MAR as well as a monotone MD Pattern (Hohl, 2007). Furthermore, the covariates in the model of the analysis have to be normally distributed (Rubin, 1987). Generation of linear discriminant functions are based on the nominal scaled dependent variable with categories  $1, \dots, g$  which are affected by missing values, continuous covariates of the dataset and the a-priori likelihoods for the categories  $1, \dots, g$ . The a-posteriori likelihoods  $p_1, \dots, p_g$  with  $p_1 + \dots + p_g = 1$  are then compared to a uniformly distributed random variable  $u$  on the interval 0 to 1. For  $u < p_1$  the missing value is imputed by category 1, for  $p_1 < u < p_1 + p_2$  by category 2 and so on. Only continuous variables can be applied to the generation of the discriminant functions.

Bootstrap methods have proved to work well for the estimation of missing values (Cohen, 1997). In small or nonparametric samples it is recommended to use respective variants of this method (Efron 2011; Pajevic & Bassler, 2003; Rubin & Schenker, 1991). The bootstrap is a useful method to estimate a valid variance for complex survey situations (Efron 2011; Shao & Sitter, 1996), in particular non-ignorable non-response (Cohen, 1997; Efron 2011; Rubin & Schenker, 1991). In the field of MD, bootstrap is based on the observed data. By resampling from the data set, random bootstrap samples are repeatedly induced (Efron & Tibshirani, 1993). Initially, the whole sample is stratified in complete and missing data and then the variability of the MD is estimated. In this context, the error variance allows a reliable approximation of the "correct bootstrap distribution". Then a MI follows (Hinkley & Davison, 1997). In non-ignorable MD situations bootstrap methods have the advantage to

guard against misspecification of the imputation model with minimal assumptions about the distribution of the data (Sapra, 2012).

In contrast to the already discussed imputation approaches, modelling approaches estimate parameters of interest without explicitly imputing MD. Most established methods are based on mixed linear models and maximum likelihood methods (Allison 2011; Efron 2011; Verbeke & Molenberghs, 2000). These advanced methods are especially indicated in the NMAR situation where the MD mechanism has to be taken into account at the estimation process (Verbeke & Molenberghs, 2000). There are different ways to partition the common distribution (Little, 1993), leading either to so called pattern mixture or selection models (Verbeke & Molenberghs, 2000). Accounting accurately for the right mechanism can often be difficult, therefore modelling approaches should be implemented with caution. They can indeed be applied when a NMAR situation is assumed, although the results remain uncertain to some extent. Therefore, it is recommended to conduct a comprehensive sensitivity analysis by applying different approaches when missing values are supposed to be NMAR (European Medicines Agency, 2010; Verbeke & Molenberghs, 2000).

Overall, the literature search suggested a superiority of MI and maximum likelihood methods over regression based approaches in case of continuous target variables (Dempster et al., 1977; Eekhout et al, 2012; Mayer, 2011; Newman, 2003). According to Allison (2009), the two particular methods of discriminant function and logistic regression imputation are proposed for categorical target variables instead of just disregard incomplete cases. However, these methods assume a monotone pattern of the data set. No direct comparisons between imputation methods and modelling approaches could be found which address the problem of MD by considering them explicitly in a statistical model. Therefore, especially in the case of NMAR, the given prioritization of modelling approaches in Figure 2 is based on the authors' subjective understanding of the methods' properties. The listed methods for the imputation of missing values have been summarized in Table 2.

**Table 2.** Imputation methods and their presumptions

Methods	Missing Data Mechanism			Scale level		Pattern	
	MCAR	MAR	NMAR	continuous	qualitative	arbitrary	monotone
<b>MCMC</b>	X	X		X		X	X
<b>FCS</b>	X	X		X	X	X	X
<b>FIML</b>	X	X	X	X	X	X	X
<b>SR</b>	X	X		X	X		X
<b>MM</b>	X	X	X	X	X	X	X
<b>EM</b>	X	X		X	X	X	X
<b>REG</b>	X	X		X			X
<b>PS</b>	X	X		X			X
<b>BB</b>	X	X	X	X	X	X	X
<b>LR</b>	X	X			X		X
<b>DFM</b>	X	X			X		X
<b>PMM</b>	X	X	X	X	X	X	X
<b>SM</b>	X	X	X	X	X	X	X

**BB** = Bayesian Bootstrap, **DFM** = Discriminant Function Method, **EM** = Estimation Maximization, **FCS** = Fully Conditional Specification, **FIML** = Full Information Maximum Likelihood, **LR** = Logistic Regression, **MAR** = Missing At Random, **MCAR** = Missing Completely At Random, **NMAR** = Not Missing At Random, **MCMC** = Markov Chain Monte Carlo, **MM** = Mixed Model, **PMM** = Pattern Mixture Model; **PS** = Propensity Score, **REG** = Regression, **SM** = Selection Model, **SR** = Stochastic Regression



## Discussion

Common MD situations and imputation strategies were reviewed. The most appropriate methods were chosen for the presented decision algorithm according to the cited literature for certain MD situations. Most SI methods were not considered since they are regarded as obsolete (Enders & Bandalos, 2001). An easily applicable algorithm (Figure 2) was suggested to assist researchers being confronted with the situation of MD in their data set in selecting an appropriate MD imputation method. Decision-making is based on MD mechanism and pattern as well as of scale level of the target variable to be imputed. NMAR is the most complicated MD mechanism to cope with. Allison (2012) described that maximum likelihood estimations were even more efficient than MI under the MAR assumption producing the same results for the same set of data. Prioritization of various available methods for specific MD situations helps to find the most suitable and, if possible, an appropriate statistical software to handle it (Mayer, Muche, & Hohl, 2012). To date this is the first decision algorithm for MD imputation methods of this complexity. Imputation of a large fraction of MD in a variable is always difficult. Then the question arises whether the chance of successful imputation is decreased. Complete case analysis then may lead to a systematic bias of study results by the loss of many observation units and the respective characteristics of other variables.

A number of articles have already proposed useful methods in several MD situations. However, finding a proper approach requires laborious and time consuming literature research. The decision algorithm was designed to efficiently support researchers in identifying appropriate imputation methods when confronted with MD. One challenge in using MD imputation procedures is that there is currently no single statistical software package in which all imputation methods are implemented (Yucel, 2011). Since not all scientists necessarily have access to commercial software packages, also non-commercial software alternatives are presented in the decision algorithm.

The stages of the algorithm provide orientation in handling various MD situations. The paths vary when the sample size is small or for different scale levels of the target variable. When applying the proposed concept, users have the possibility to impute MD in single-item instruments in several situations with various imputation methods.

Imputation methods are listed following the three stages MD mechanism, scale level and MD pattern. Methods were listed in order of their appropriateness as described in the cited literature. The suggested statistical software packages and references for each method are listed at the upper and lower left corners of each method block. Deciding between systematic missing (NMAR) and not (MCAR, MAR) is challenging and will surely demand certain understanding and consideration. However, upon choosing the initial paths, the user should obtain results efficiently via suggested statistical software packages.

Completeness of the presented methods cannot be guaranteed. The suggested prioritization is mainly based on findings in cited references, which compared different methods for imputing MD. Results of future studies comparing different imputation strategies may require a revision of this prioritization. A crude understanding of the methods described in the algorithm is advantageous, therefore a study of the cited references in the algorithm and to consult an expert is recommended. There is still no universal recommendation in relevant guidelines (European Medicines Agency, 2010) and the literature with respect to the percentage of MD up to which an imputation is suggested. Further analysis is needed to prove the applicability established methods for different amounts of MD. Furthermore, every

method given in the algorithm has its limitations, for example the use of selection models in the NMAR situation: if the assumptions are satisfied, the selection model can virtually eliminate bias caused by NMAR data, but already modest correlations among the predictor variables and moderate deviation from normality assumption may produce severely biased estimates. Some limitations might will be handled by future developments.

Overall, our extensive literature search revealed that there are numerous approaches which address the problem of MD in medical research. However, it has obviously not been possible yet to arrange them reasonably to have a universal guidance for their application in specific MD situations. This points out again the importance of making all efforts to generally prevent MD and to find ways to systemize the available methods.

#### Acknowledgement

The authors thank Prof. Dr. Dietrich Rothenbacher and Prof. Dr. Rainer Muche for their helpful advice and critical comments on drafts of the manuscript. This research was supported by German Research Foundation Grant awarded to Hatice Uenal and by the German Federal Ministry of Research and Education Grant awarded to Jean-Baptist du Prel.

#### References

1. Allison, P.D. **Missing data.** in: Millsap, R.E. and Maydeu-Olivares, A. (eds.) "The SAGE Handbook of Quantitative Research in Psychology. 1st ed. London", Thousand Oaks, New Dehli, Singapore: Sage Publications, 2009, pp. 72-89
2. Allison, P.D. **Handling Missing Data by Maximum Likelihood.** SAS Global Forum, 2012
3. Baraldi, A.N. and Enders, C.K. **An introduction to modern missing data analyses.** Journal of School Psychology, Vol. 48, 2010, pp. 5-37
4. Cohen, M.P. **The Bayesian bootstrap and multiple imputation for unequal probability sample designs.** Proceedings of the Survey Research Methods Section, American Statistical Association, 1997, pp. 635-638
5. Dempster, A. P., Laird, N. M., and Rubin, D. B. **Maximum Likelihood for Incomplete Data via the EM Algorithm.** Journal of the Royal Statistical Society, Vol. 39, 1977, pp. 1-38
6. Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C. and Heymans, M. W. **Missing data: a systematic review of how they are reported and handled.** Epidemiology, Vol. 23, 2012, pp. 729-732
7. Efron, B. **The bootstrap and Markov chain Monte Carlo.** Journal of Biopharmaceutical Statistics, Vol. 21, 2011, pp. 1052-1062
8. Efron, B. and Tibshirani, R. J. **An Introduction to the Bootstrap.** 1st ed. New York, NY: Chapman and Hall, 1993
9. Enders, C.K. **Applied Missing Data Analysis (Methodology in the Social Sciences).** 1st ed. New York, London: Guilford Press, 2010
10. Enders, C.K. and Bandalos, D.L. **The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models.** Structural Equation Modelling: A Multidisciplinary Journal, Vol. 8, 2010, pp. 430-457
11. European Medicines Agency. **Guideline on Missing Data in Confirmatory Clinical Trials.** European Medicines Agency, London, 2010

12. He, Y. **Missing data analysis using multiple imputation: getting to the heart of the matter.** *Circulation Cardiovascular Quality and Outcomes*, Vol. 3, 2010, pp. 98-105
13. Hinkley, D. V. and Davison, A.C. **Bootstrap Methods and their Application.** 1st ed. Cambridge, England: Cambridge University Press, Publishers, 1997
14. Hohl, K. **Handling missing data: imputation methods for missing categorical values in clinical datasets.** PhD Thesis 2007, University of Ulm (in German); [http://vts.uni-ulm.de/docs/2007/6027/vts\\_6027\\_8103.pdf](http://vts.uni-ulm.de/docs/2007/6027/vts_6027_8103.pdf). Accessed June 2014
15. Lavori, P. W., Dawson, R. and Shera, D. **A multiple imputation strategy for clinical trials with truncation of patient data.** *Statistics in Medicine*, Vol. 14, 1995, pp. 1913-1925
16. Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T. and Stern, H. **The prevention and treatment of missing data in clinical trials.** *New England Journal of Medicine*, Vol. 367, 2012, pp. 1355-1360
17. Little, R. J. A. **Regression with missing X's: a review.** *Journal of the American Statistical Association*, Vol. 87, 1992, pp. 1227-1237
18. Little, R. J. A. **Pattern mixture models for multivariate incomplete data.** *Journal of the American Statistical Association*, Vol. 88, 1993, pp. 125-134
19. Little, R. J. A. and Rubin, D. B. **Statistical Analysis with Missing Data.** 2nd ed. New York, NY: John Wiley & Sons, 2002
20. Mayer, B. **Missing values in clinical longitudinal studies – handling of drop-outs.** PhD Thesis 2011, University of Ulm. (in German); [http://vts.uni-ulm.de/docs/2011/7633/vts\\_7633\\_10939.pdf](http://vts.uni-ulm.de/docs/2011/7633/vts_7633_10939.pdf). Accessed June 2014
21. Mayer, B., Muche, R. and Hohl, K. **Software for the Handling and Imputation of Missing Data – An Overview.** *Journal of Clinical Trials*, Vol. 2, 2012, p. 103
22. Newman, D. A. **Longitudinal Modelling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques.** *Organizational Research Methods*, Vol. 6, 2003, pp. 328-362
23. Pajevic, S. and Basser, P. J. **Parametric and Non-Parametric Statistical Analysis of DT-MRI Data.** *Journal of Magnetic Resonance*, Vol. 161, 2003, pp. 1-14
24. Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V. and Solenberger, P. **A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of regression Models.** *Survey Methodology*, Vol. 27, 2001, pp. 85-95
25. Rubin, D. B. **Inference and missing data.** *Biometrika*, Vol. 63, 1976, pp. 581-592
26. Rubin, D. B. **Multiple Imputation for Nonresponse in Surveys.** 1st ed. New York, NY: John Wiley & Sons, 1987
27. Rubin, D.B. and Schenker, N. **Multiple Imputation in health-care databases: An overview and some applications.** *Statistics in Medicine*, Vol. 10, 1991, pp. 585-598
28. Sapra, S. **Robust bootstrap multiple imputation with missing data and outliers.** <http://ubuntuone.com/75CiQjUOlglXFG2utL9WWhr>. Accessed June 2014
29. Schafer, J. L. and Graham, J. W. **Missing data: Our view of the state of the art.** *Psychological Methods*, Vol. 7, 2007, pp. 147-177

30. Sentas, P., Angelis, L. and Stamelos, I. **Multiple Logistic Regression as Imputation Method Applied on Software Effort Prediction.** Department of Informatics, Aristotle University of Thessaloniki. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.196.1080>. Accessed June 2014
31. Shao, J. and Sitter, R. R. **Bootstrap for Imputed Survey Data.** Journal of the American Statistical Association, Vol. 91, 1996, pp. 1278-1288
32. Tanner, M. A. and Wong, W. H. **The Calculation of Posterior Distributions by Data Augmentation.** Journal of the American Statistical Association, Vol. 82, 1987, pp. 528-540
33. Vach, W. and Blettner, M. **Biased Estimation of the Odds Ratio in Case-Control Studies due to the use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables.** American Journal of Epidemiology, Vol. 134, 1991, pp. 895-907
34. Verbeke, G. and Molenberghs, G. **Linear mixed models for longitudinal data.** New York, NY: Springer Press, 2000
35. Yucel, R. M. **State of the Multiple Imputation Software.** Journal of Statistical Software, Vol. 45, 2011, pp. 1-7