Evaluating Imputation Methods for Missing Data in a MCI Dataset

Alba Gómez-Valadés Batanero(✉), Mariano Rincón Zamorano, Rafael Martínez Tomás, and Juan Guerrero Martín

Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain
albagvb@dia.uned.es

Abstract. Missing data is a recurrent problem in experimental studies, mostly in clinical and sociodemographic longitudinal studies due to the dropout and the negative of some subjects to answer or perform some tests. To address this problem different strategies have been designed to deal with missing values, but incorrect treatment of missing data can result in the database being biased in one or more parameters, compromising the viability of the database and future studies. To solve this problem different imputation techniques have been developed over the last decades. However, there are no regulations or clear guidelines to deal with these situations. In this study, we will analyze and impute a real, incomplete database for the early detection of MCI, where the loss of values on 3 main variables is strongly correlated with the years of studies. The imputation will follow two strategies: assuming that those people would have got a bad scoring if they had taken the test, defining a ceiling score, and a multiple imputation by fully conditional specification. To determine if any kind of bias in mean and variance has been introduced during the imputation, the original database was compared with the imputed databases. Taking a p-value = 0.1 threshold, the database imputed by the multiple imputation method is the one that best preserved the information of the original database, making it the more appropriate imputation method for this MCI database.

Keywords: Missing data · Imputation · Multiple imputation

1 Introduction

Missing data is a persistent problem mostly in clinical and sociodemographic studies [1,2], mostly in longitudinal studies, where subjects can drop from the study due to illness, ilocation, negative to follow in the study, or death. Between those that keep with the study, there is always the possibility that subjects refuse to do some of the tests [1]. Missing data may seriously compromise inferences from randomized clinical trials, especially if missingness is not at random and if missing data are not handled appropriately, causing potential bias that distorts parameters and relationships between variables which can cause misleading results and conclusions [2–4].

J. M. Ferrández Vicente et al. (Eds.): IWINAC 2022, LNCS 13258, pp. 446–454, 2022.
https://doi.org/10.1007/978-3-031-06242-1_44

To overcome the problems caused by missing data, there has been extensive development of statistical models and software for imputing the data, or by directly analyzing data with missing values [5]. However, in order to select the best option for a dataset, as a first step is necessary to analyze the causes and the distribution of the missing data. Data may be missing due to one of following three reasons: Missing completely at Random (MCAR), Missing at Random (MAR), and Missing Not At Random (MNAR) [6,7]. If the mechanism causing the missing data does not depend on observed or unobserved variables, the data is called MCAR. The deletion of records with missing data does not introduce any bias since the reduced dataset is representative of the original dataset, being the only problem the reduction of the sample size [2,3,8]. If the missing data only depends on the observed variables it is called MAR, where is possible to predict missing values given the existing data [2,3]. However, the deletion of the cases with missing values will generate a biased subdataset. If the distribution of missing data depends on unobserved variables, it is called MNAR. This situation is the most difficult to be detected and to deal with since the relationship is not contained in the database itself and is therefore unknown, and usually causes bias in the subdataset whether the records are imputed or deleted [2,3,9]. Missing data can also be classified as ignorable and non-ignorable. Missing data is ignorable if the probability of observing a data item is independent of the value of that item, and it corresponds with the MCAR and MAR scenarios. If the probability of observing a data item is dependent on the value of that item, the missing data is non-ignorable, and it corresponds with the MNAR scenario [10].

To cope with this missing data problem, especially in scenarios where deletion of missing data is not recommended due to strong dataset reduction and/or introduction of bias, different types of strategies and techniques to estimate the effect on the missing data have been developed [8,9]. The simplest method is the complete case analysis, where all observations with missing data are discarded, but is prone to introduce bias in no-MCAR scenarios [9]. When deletion of records is not considered an adequate option, the data imputation by using the observed data and employing a series of rules is used. The most basic imputation method is to complete the missing values with the mean, median, or mode. Like complete case analysis, this strategy is prone to introduce bias [11], but it can be useful if just a handful of scores are missing. Imputations based on statistical approaches, like logistic regression or k-nearest neighborhoods, have been proposed, but while they may produce better solutions, they may also create a distorted dataset if the assumptions are broken [11]. Finally, multiple imputation is a widely used method because it is a simple and powerful strategy to impute missing data [13]. Currently, two major iterative methods are used for doing multiple imputation: joint modeling (JM) and fully conditional specification (FCS) [11,12]. JM is based on parametric statistical theory and leads to imputation procedures whose statistical properties are known [12]. However, it assumes normality and linearity, and also it is often difficult to realistically specify a joint model, potentially leading to bias [11]. FCS generates imputations by iterating over the conditional densities on a variable-by-variable basis, given an starting point [11,12]. However, depending on the quantity and the distribution of the missing data, sometimes more simple methods can achieve results as good as multiple imputation. In this context, it is preferable to first make a statistical analysis of the missing data distribution, and the possible relationships between variables before selecting a strategy.

In this paper, we will study the missing value distribution and the problem of the imputation of missing data in the database of a longitudinal study of the prevalence and evolution of MCI in a group of monolingual Spanish subjects. Therefore, this database presents the characteristics indicated above and makes it necessary to use these imputation techniques so that the studies carried out with it are valid and informative. In order to select the most appropriate method to avoid losing information or altering the information in future analyses, we will compare two different imputation methodologies.

This study is part of a broader project focused on the comparison of two databases obtained from subjects with different sociodemographic backgrounds, in order to identify the most discriminating variables, as well as those machine learning systems with the best performance in predicting the diagnosis, with the idea of creating an expert system that could help in diagnosis and save time for physicians in the evaluation of certain tests that are currently performed by hand.

2 Methodology

2.1 Database Description

The database consists of a sample from a large ongoing longitudinal study with the aim of discerning what types of tests, subtests, or sociodemographic variables had the greatest influence when making an early diagnosis of MCI and to determine the prevalence and stability of MCI in the Autonomous Community of Madrid (Spain) [14,15]. As it is shown in Table 1, the database is composed of 947 cases of Spanish monolingual subjects, with 540 cases with missing values. The mean age is nearly 71 years and a standard deviation of 6’31 years. The education, measured in schooling years, has a mean of 11’5 years of schooling, with a standard deviation of 6’45 years. The database is split up to 4 evaluations, with 428 cases on first evaluation, 319 on second evaluation, 171 on third evaluation, 18 on forth evaluation and 11 of unknown evaluation. Each evaluation corresponds to the completion of the entire test battery by the same subject, with an average difference of 1 year between evaluations.

The database used in this study is composed of a total of 32 variables, described in more detail in Table 2, and grouped in identifiers and database metadata, sociodemographic variables, psychologic tests, diagnosis, and independent blocks of screening variables. The screening variables are in turn made up of the groups of MMSE, Verbal Fluency, Graphic tests, Ideomotor Tests, Trail Marking Tests, Rey Figure’s tests, and the TAVEC, and integrates the data used to assess if a patient is healthy or MCI, and are the variables in which we will

Table 1. Distribution of the Spanish dataset cases and variables.

Total cases (without missing values)	947 (407)
Nº assessments	4
Total of variables	32
Age (Mean/Standard Deviation)	70,88/6,31
Years of schooling (Mean/Standard Deviation)	11,57/6,45

focus this study, along with the sociodemographic variables and the relationships between the two groups.

2.2 Missing Values Analysis

As a first step, an analysis was performed to determine the distribution of the missing values, identify those variables that concentrate most of them, and identify possible relationships between data and the missing values. A distribution analysis of the missing values across rows was performed in order to detect nearly empty rows that must be discarded. The reason is that more than 40% of missing data makes it impossible to perform a feasible imputation on those rows.

Next, a more in deep analysis of the possible relationship between the variables with the most missing values and the sociodemographic variables of sex, age and schooling was carried out to identify possible links between them, which will be taken into account during the allocation process. The performance of a subject on a test is related to their sociodemographic context, and it is the reason after the fact that the tests are calibrated due to those variables, so it was possible that the missing values were correlated to those variables.

2.3 Imputation Strategy

With the data obtained in the analysis, the database was imputed using two different strategies.

The first strategy consists of imputing the dataset using an ad hoc value. The variables with less than 5 missing values will be imputed with the mode. The mode was chosen due to the low number of missing values and the distribution of the data. In the 3 variables in which more values were missing (Trail Marking Test B (TMTB), Rey Figure's score (RFS) and Rey Figure's time (RFT)), a ceiling value was defined. To perform these tests is necessary a certain degree of abstraction. In the case of the TMTB, it is necessary to know both the alphabet and the numbers, and to perform an abstraction in order to be able to alternate between letters and numbers in the correct order. In the case of the Rey Figure, the subject have to copy a completely abstract complex figure in the shortest possible time. Subjects with a low level of education have problems to perform the abstraction necessary to do these tests properly, so their score tends to be worse than subjects with the same mental condition but with a higher education

Table 2. Description of the variables used in the database.

Group	Variables name	General description
Identifiers	Nº of the column, subject ID evaluation	Allow to identify each case individually
Sociodemographic	Sex, years of schooling age	Sociodemographical variables
Psychiatric tests	Blessed Yesavage	Evaluation of the mood of the subject, to exclude those unsuitable for the study
MMSE	MEC	Spanish version of the Mini-mental test
Verbal fluency	Phonetic fluency Animal Clothes Plants Vehicles Semantic Fluency	Maximum number of correct words according to a certain rule said in 1 min
Graphic tests	Peak_Loops Constructive Praxias	To make a copy of a series of simple drawings
Ideomotor tests	Ideomotor Ideopatic Ideototal	Mimic a verbally described action
Trail Marking Tests	Trail Marking Test A Trail Marking Test B	Join with a line a series of numbers or/and letters in ascending order
Rey Figure	Rey Figure Score Rey Figure Time	To make a copy the complex Rey Figure
TAVEC (Test de Aprendizaje Verbal España-Complutense)	Total List A Total List B Short-term recall Short-term recall with clues Long-term recall Long-term recall with clues Recognition	Evaluation of short-term and long-term memory with and without clues
Diagnosis	Profiles	Subject diagnosis: healthy or MCI

level. Thus, it was assumed that these missing values were due to the inability of those subjects to perform the tests. Therefore, those subjects would have obtained an bad score on those tests. For the TMTB we followed the psychologist criteria, which defines a ceiling score of 300 s for the TMTB when the subject struggles for too long in this test. For the RFS and RFT scores we had no previous ceiling scores, so we obtained it from the data. We analyzed the scores on the range (from 0 to 5 years of schooling) in which the missing values were concentrated, and the worst most common scores were selected. Using the overall worst score was discarded due to the presence of clear outliers for the RFT, and because in the RFS, a discrete variable, the occurrence of very low scores was exceptionally uncommon. Having this into account, we defined a value of 15 for the RFS, and of 210 for the RFT.

The second strategy was to impute the database using multiple imputation with the FCS strategy, in which we impute the dataset 10 times using the FCS algorithm. Then we grouped the results to generate the final imputed database. All variables were used during the multiple imputation.

Finally, mean and variance between original and imputed databases were compared to determine if any type of bias was introduced and, therefore, which imputed database will be more recommendable for future analysis. For the mean comparisons the t-test was used, and for the variance comparison, the Levene test was used. Python programming language was used for this study, using the scikit-learn package for certain parts of the analysis and imputation.

3 Results

The date variables of evaluation date and birth-date were removed they are completely redundant with the Age sociodemographic variable.

In the missing value analysis per variable, we found that a big percentage of the variables had a similar quantity of missing values, including sociodemographic variables of sex, age, and schooling, and 3 variables with a number of missing values much higher than the others, TMTB, RFS and RFT. The analysis of the missing values by rows showed the presence of records with more than 50% of missing values. Those rows were discarded since the great percentage made unreliable a correct imputation. In the distribution analysis of missing values that were then carried out after that, the missing values on all variables were greatly reduced, with some variables achieving 0 missing values, with the TMTB, RFS, and RFT variables maintaining a much higher percentage of missing values than the rest of variables.

The distribution analysis of missing values of TMTB, RFS, and RFT showed a great correlation between low education years and missing values. The comparative analysis of the comparison between the global dataset and the subset corresponding to the schooling years between 0 and 5 showed that there are differences in both mean and variance in all variables except one. Since there is a natural bias between subjects with high levels of education and those with low levels of education, the subset focused on years of education from 0 to 5 years will be used in both cases to make the comparison between the original and the imputed database. By using the subtest, we transform the MAR problem into a MCAR, where the distribution and relationship between the original database and the imputed ones can be comparable since the bias caused by years of education is eliminated.

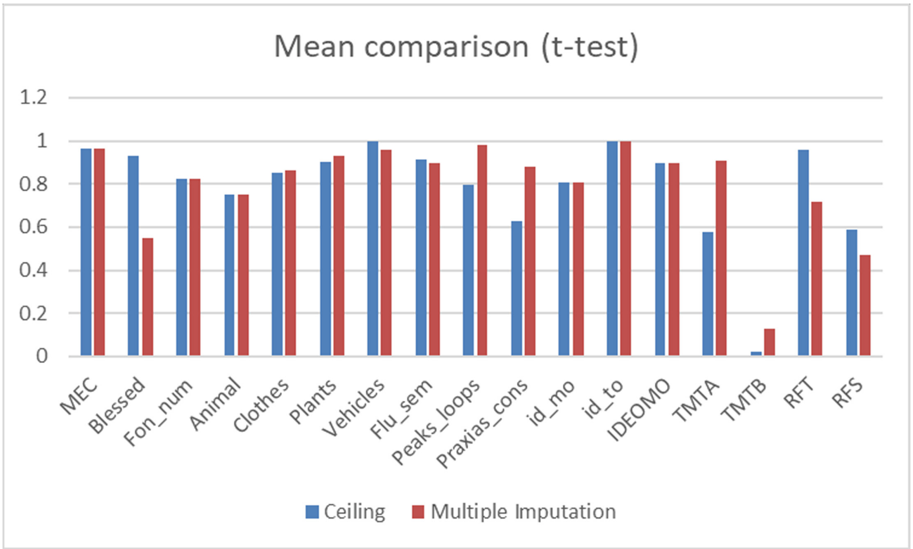

Fig. 1. P-values obtained in the comparison of means for ceiling imputation and multiple imputation.

In (Fig. 1) is shown the mean comparison on both imputed databases in all imputed variables. Taken a p-value = 0,1 as the threshold, no significant differences on the mean were founded between the original and the imputed database except on the TMTB, in which p-value = 0,02. It is also noticeable how multiple imputation presents in general better results than ceiling imputation, except in the variables of RFT, RFS and Blessed. On the comparison between variances showed in (Fig. 2), using a p-value = 0,1 as threshold, both methods do not introduce any significant bias in the data. However, it is important to note that multiple imputation introduces less bias than ceiling imputation, except in the Blessed variable. Taking the data obtained from the mean and variance together, we obtain that multiple imputation is a more suitable method to impute the database, making this imputation more suitable for future analysis on the database.

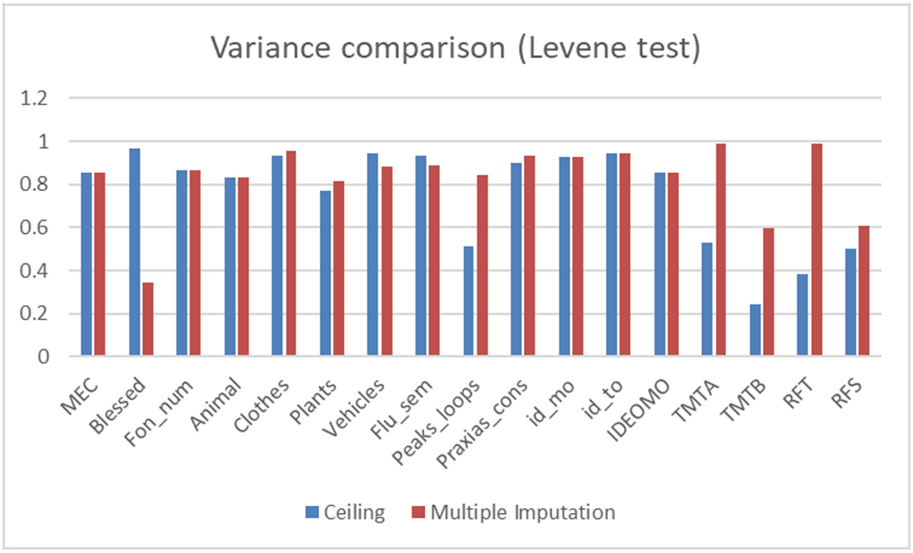

Fig. 2. P-values obtained in the comparison of variances for ceiling imputation and multiple imputation.

4 Conclusions

In this work we used different imputation method to improve the quality of the MCI dataset, allowing to increase the number of usable cases from 407 to 910 and to work with a no-biased database. The missing values were centered on 3 variables of high interest in the early detection of the MCI, due to its high discrimination and because they are some of the most common tests in MCI screening batteries. This make not recommendable to remove those scores. However, maintaining those variables by eliminating missing records would have generated a bias. The imputation allows these tests to be maintained by losing as little information as possible, allowing the use of this database in future analyses keeping the highest percentage of information. To achieve this, two imputation methods carried out in a real MCI database with missing values. The results showed that, although the ceiling imputation can be useful when the values are lost in a MAR situation and the correlation between values is clear, multiple imputation shows to be completely unbiased in all analyzed aspects. A more in deep analysis of the Blessed variable, the only variable in which multiple imputation obtain a worst score in both mean and variance as ceiling imputation was left as future work.

References

1. Nguyen, C.D., Carlin, J.B., Lee, K.J.: Model checking in multiple imputation: an overview and case study. Emerg. Themes Epidemiol. **14**(1), 8 (2017)

2. Sterne, J.A.C.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ **338**, b2393 (2009)
3. Jakobsen, J.C., Gluud, C., Wetterslev, J., Winkel, P.: When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med. Res. Methodol. **17**(1), 162 (2017)
4. Groenwold, R.H.H., Moons, K.G.M., Vandenbroucke, J.P.: Randomized trials with missing outcome data: how to analyze and what to report. Can. Med. Assoc. J. **186**(15), 1153–1157 (2014)
5. Hughes, R.A., Heron, J., Sterne, J.A.C., Tilling, K.: Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int. J. Epidemiol. **48**(4), 1294–1304 (2019)
6. Rubin, D.R.: Inference and missing data. Biometrika **63**(3), 581–590 (1976)
7. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York (1987)
8. Dziura, J.D., Post, L.A., Zhao, Q., Fu, Z., Peduzzi, P.: Strategies for dealing with Missing data in clinical trials: from design to analysis. Yale J. Biol. Med. **86**, 343–8358 (2013)
9. Choi, J., Dekkers, O.M., le Cessie, S.: A comparison of different methods to handle missing data in the context of propensity score analysis. Eur. J. Epidemiol. **34**(1), 23–36 (2018). https://doi.org/10.1007/s10654-018-0447-z
10. Marlin, B.M., Roweis, S.T., Zemel, R.S.: Unsupervised Learning with Non-Ignorable Missing. AISTATS (2005)
11. Liu, Y., De, A.: Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. Int. J. Stat. Med. Res. **4**(3), 287–295 (2019)
12. van Buuren, S.: Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. **16**(3), 219–242 (2007)
13. Murray, J.S.: Multiple imputation: a review of practical and theoretical findings. Stat. Sci. **33**(2), 142–159 (2018)
14. Peraita, H., García-Herranz, S., Díaz-Mardomingo, M.C.: Evolution of specific cognitive subprofiles of mild cognitive impairment in a three-year longitudinal study. Curr. Aging Sci. **4**, 171–182 (2011)
15. García-Herranz, S., Díaz-Mardomingo, M.C., Venero, C., Peraita, H.: Accuracy of verbal fluency tests in the discrimination of mild cognitive impairment and probable Alzheimer's disease in older Spanish monolingual individuals. Neuropsychol. Dev. Cogn. Section B, Aging, Neuropsychol. Cogn. **27**(6), 826–840 (2020)