A framework for testing different imputation methods for tabular datasets

Tabea Kossen[1,2*¤], Michelle Livne[1,2], Vince I Madai[1,2], Ivana Galinovic[3], Dietmar Frey[1,2], Jochen B Fiebach[3]

1 Department of Neurosurgery, Charité - Universitätsmedizin Berlin, Berlin, Germany

2 Berlin Institute of Health, Berlin, Germany

3 Center for Stroke Research Berlin (CSB), Charité - Universitätsmedizin Berlin, Berlin, Germany

¤Current Address: PREDICTioN 2020, Klinik für Neurochirurgie CCM - Forschung, Charité-Universitätsmedizin, Charitéplatz 1, 10117, Berlin, Germany

* tabea.kossen@charite.de (TK)

Abstract

Background and purpose

Handling missing values is a prevalent challenge in the analysis of clinical data. The rise of data-driven models demands an efficient use of the available data. Methods to impute missing values are thus crucial. Here, we developed a publicly available framework to test different imputation methods and compared their impact in a typical stroke clinical dataset as a use case.

Methods

A clinical dataset based on the 1000Plus stroke study with 380 completed-entries patients was used. 13 common clinical parameters including numerical and categorical values were selected. Missing values in a missing-at-random (MAR) and missing-completely-at-random (MCAR) fashion from 0% to 60% were simulated and consequently imputed using the mean, hot-deck, multiple imputation by chained equations, expectation maximization method and listwise deletion. The performance was assessed by the root mean squared error, the absolute bias and the performance of a linear model for discharge mRS prediction.

Results

Listwise deletion was the worst performing method and started to be significantly worse than any imputation method from 2% (MAR) and 3% (MCAR) missing values on. The underlying missing value mechanism seemed to have a crucial influence on the identified best performing imputation method. Consequently no single imputation method outperformed all others. A significant performance drop of the linear model started from 11% (MAR+MCAR) and 18% (MCAR) missing values.

Conclusions

In the presented case study of a typical clinical stroke dataset we confirmed that listwise deletion should be avoided for dealing with missing values. Our findings indicate that the underlying missing value mechanism and other dataset characteristics strongly influence the best choice of imputation method. For future studies with similar data structure, we thus suggest to use the developed framework in this study to select the most suitable imputation method for a given dataset prior to analysis.

Introduction

Missing values are a prevalent challenge in the analysis of clinical data [1–3]. Validated guidelines for handling missing data are more important now than ever with the rise of data-driven applications [4–8]. Here, an efficient utilization of the data is crucial in the limited settings of usually fairly small medical datasets [9]. Additionally, while deletion of patient entries with missing values (listwise deletion) is a common practice, it can lead to biased results and is therefore highly discouraged [1, 10, 11].

An alternative approach is to apply imputation methods on the missing data [12, 13]. Imputation methods allow replacing missing values with substituted values that estimate the true underlying value. Most commonly used methods in clinical datasets include simple imputation methods like mean imputation and hot-deck imputation ("sampling") as well as more complex algorithms like multiple imputation by chained equations (MICE) [14] and expectation maximization (EM) using multiple imputation [15].

However, there is a controversy regarding which imputation method should be used [12, 16, 17]. More importantly, only few studies exist which assessed imputation methods in the medical field [13, 16–18]. This is also true in the stroke field. While Young-Saver et al. investigated the imputation of stroke outcome data, to date no study has compared and validated imputation methods for a typical clinical stroke dataset as a whole [19].

The objectives of this work were thus to 1) develop a publicly available framework (`https://github.com/tabeak/missing-value-analysis`) which can identify the best imputation method for a given tabular dataset and 2) to compare different imputation method for handling missing data in clinical stroke dataset as a use case for the framework. When comparing the different imputation methods, we assessed both how much the imputed values differed from the ground truth as well as how the different imputation methods influenced the performance of a data-driven predictive model.

Materials and methods

Patients

A clinical dataset based on the 1000Plus stroke study with 380 completed-entries of acute stroke patients was used [20]. The study was approved by the institutional review board of the Charité Universitätsmedizin Berlin. All patients gave their written informed consent. Because of the sensitive nature of the data collected for this study, requests to access the dataset from qualified researchers trained in human subject confidentiality protocols may be sent to the institutional ethics commitee of Charité Universitätsmedizin Berlin.

Data analysis

13 common clinical parameters were analyzed. Among them, seven were numerical: hours-to-MRI, age, pre-stroke mRS, acute National Institutes of Health Stroke Scale

(NIHSS), discharge NIHSS, discharge mRS and discharge Trial-of-ORG-10172-in-Acute-Stroke-Treatment (TOAST). Six were categorical: sex, treatment with tissue plasminogen activator (tPA), occlusion, hyperlipidemia, diabetes and hypertonia.

Missing values from 0% to 60% were simulated following two different cases of missing values: missing-at-random (MAR) and missing-completely-at-random (MCAR). MAR means that the probability for a value missing depends on same values of other observed variables. MCAR, on the other hand, describes the scenario that values are missing completely at random. In contrast to MAR, there is no systematic reason for a missing value and the probability for a value missing is the same for each value. Performance was estimated for different imputation methods in two fashions: 1) Error assessment using RMSE and absolute bias and 2) Performance assessment of stroke discharge mRS. The imputation methods included 1) mean imputation, 2) hot-deck imputation, 3) MICE and 4) multiple imputation by EM and 5) listwise deletion.

Error assessment

For the error assessment analysis, we chose two common measures for evaluating imputation methods, RMSE and absolute bias [21]. In this analysis, the parameters were split into numerical and categorical. For numerical parameters, the RMSE of the normalized data is defined according to:

$$\mathrm{RMSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(Y_i - \hat{Y}_i)^2}, \tag{1}$$

where n is the number of imputed samples, $\hat{Y}_i$ the estimated sample value and Y_i the true value. For categorical parameters, the RMSE corresponds to the percentage of misclassified values:

$$\%\ \text{of misclassified samples} = \frac{\text{number of misclassified samples}}{\text{total number of samples}}, \tag{2}$$

As a second error assessment, the mean absolute bias was calculated. It is defined as:

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\sum_{j=1}^{m}\hat{Y}_j\right)_i - Y_i\right|, \tag{3}$$

where m is the number of iterations for each imputation, i.e. how often the value was imputed. The absolute bias was then averaged over all imputed samples n.

Both RMSE as well as the absolute bias was assessed for each variable at a time and then averaged for each parameter-type (i.e. numerical vs. categorical).

Predictive model analysis

The second part of the analysis included the incorporation of a supervised predictive model, the generalized linear model (GLM). We constructed a model predicting the modified Rankin Scale (mRS) at discharge, which is a measure of the early clinical outcome after stroke. It can take values from 0 (no symptoms) to 6 (death). The mRS was split into 0-2 (good outcome) and 3-6 (bad outcome) [4, 22]. Other discharge parameters (discharge NIHSS and discharge TOAST) were excluded for this analysis to maintain the integrity of the model. This predictive modelling framework is used as a standard method for this use case [23–25]. As the MAR mechanism deletes parameter values with respect to other parameter's values, the mechanism could only be simulated for up to 9% missing values. Therefore, missing values from 10% to 60% are deleted combining the MAR mechanism with the MCAR mechanism. For simplicity, this mechanism is hence termed "MAR+MCAR" throughout this work.

Receiver operating characteristics (ROC) analysis was applied to assess the model performance and the imputation methods were then compared based on the area under the curve (AUC). Further, the methodologies were compared to listwise deletion. Listwise deletion removes every patient that has at least one missing value and thus, reduces the size of the dataset. Due to this limitation, the GLM could only be computed for up to 10% missing values for this sub-analysis.

Statistical analysis

Both the error and the predictive model analyses were repeated 100 times for randomly simulated missing values. For the RMSE, each imputation method was compared to each of the other methods using the paired Wilcoxon signed-rank test for each respective percentage value. The same analysis was conducted for performance assessment of the predictive model. To calculate the mean absolute bias, the 100 repeated imputations are averaged and then compared to the true value. This results in one value per percentage. Thus, a statistical analysis cannot be performed for each percentage.

Finally, the threshold of the percentage of missing values to significantly impair performance was determined for the different imputation methods: The performance of the completed-entry dataset (0% missing values) was compared to the performance for each percentage of missing value using the paired Wilcoxon signed-rank test. The threshold was identified when the performance drop was found to be significant according to the standard value of $p < 0.05$.

Results

Clinical data

380 acute stroke patients had complete entries for the 13 clinical parameters. The median age was 72 and the median NIHSS score was 3. The distribution of the clinical parameters is given in Table 1.

Error assessment

Figs 1 to 4 show the RMSE and absolute bias for increasing MAR and MCAR missing values. The figures are separated into the averaged and the accumulated values of both numerical (Figs 1 and 4) and categorical data (Figs 2 and 4). While the accumulated error increases with larger percentages of missing values, the averaged RMSE and absolute bias remains relatively steady. No single imputation method consistently outperformed all other methods.

The RMSE is shown in Figs 1 and 2. For the numerical data MICE and mean imputation show the lowest error rate for MAR missing values and mean imputation for MCAR missing values (Fig 1). For the categorical data the lowest percentage of misclassified samples could be observed for mean imputation (Fig 2). For MAR missing values the mean imputation appears less steady and stable compared to the MCAR

Table 1. Original distribution of clinical parameters.

Clinical parameter	Value
median hours-to-MRI (IQR)	11 (15)
median age (IQR)	72 (15)
median mRS prestroke (IQR)	0 (1)
median acute NIHSS (IQR)	3 (4.25)
median discharge NIHSS (IQR)	1 (3)
median discharge mRS (IQR)	1 (2)
median discharge TOAST (IQR)	1 (1)
females (%)/ males (%)	247 (65) / 133 (35)
tPA treatment yes / no (%)	93 (24.47) / 287 (75.53)
occlusion yes / no (%)	219 (57.63) / 161 (42.37)
hyperlipidemia yes / no (%)	224 (58.95) / 156 (41.05)
diabetes yes / no (%)	286 (75.26) / 94 (24.74)
hypertonia yes / no (%)	309 (81.32) / 71 (18.68)

Median values and IQR of the numerical clinical parameters hours-to-MRI, age, mRS pre-stroke, acute NIHSS, discharge NIHSS, discharge mRS and discharge TOAST. Total amount of patients and percentages of the categorical parameters sex, treatment with tPA, occlusion, hyperlipidemia, diabetes and hypertonia. (Abbreviations: IQR = interquartile range, MRI = magnetic resonance imaging, mRS = modified Rankin scale, NIHSS = National Institutes of Health Stroke Scale, TOAST = Trial-of-ORG-10172-in-Acute-Stroke-Treatment, tPA = tissue plasminogen activator).

Fig 1. Mean RMSE for increasing percentages of missing data for numerical parameters. Mean RMSE for increasing missing values from 0% to 60% using mean imputation (blue), hot-deck imputation (olive), MICE (purple) and EM (red) for numerical data. (A) describes the accumulated average RMSE if missing values are generated in a MAR fashion, (B) in a MCAR fashion. (C) shows the average RMSE if missing values are generated using the MAR mechanism and (D) the MCAR mechanism. (Abbreviations: EM = expectation maximization, MAR = missing-at-random, MCAR = missing-completely-at-random, MICE = multiple imputation by chained equations, RMSE = root mean squared error)

Fig 2. Mean percentage of misclassified samples for increasing percentages of missing data for categorical parameters. Mean percentage of misclassified samples for increasing missing values from 0% to 60% using mean imputation (blue), hot-deck imputation (olive), MICE (purple) and EM (red) for categorical data. (A) describes the accumulated average percentage of misclassified samples if missing values are generated in a MAR fashion, (B) in a MCAR fashion. (C) shows the average percentage of misclassified samples if missing values are generated using the MAR mechanism and (D) the MCAR mechanism. (Abbreviations: EM = expectation maximization, MAR = missing-at-random, MCAR = missing-completely-at-random, MICE = multiple imputation by chained equations)

missing values.

Both mean imputation and MICE show a significantly lower RMSE than hot-deck and EM for numerical data in the MAR type-case ($p < 0.05$). For MCAR missing values as well as MAR missing values on categorical data mean imputation in terms of

Fig 3. Mean absolute bias for increasing percentages of missing data for numerical parameters. Mean absolute bias for increasing missing values from 0% to 60% using mean imputation (blue), hot-deck imputation (olive), MICE (purple) and EM (red) for numerical data. (A) describes the accumulated average bias if missing values are generated in a MAR fashion, (B) in a MCAR fashion. (C) shows the average bias if missing values are generated using the MAR mechanism and (D) the MCAR mechanism. Note that the error margin in (C) and (D) corresponds to the standard deviation of the samples estimates and not the bias. (Abbreviations: EM = expectation maximization, MAR = missing-at-random, MCAR = missing-completely-at-random, MICE = multiple imputation by chained equations)

Fig 4. Mean absolute bias for increasing percentages of missing data for categorical parameters. Mean absolute bias for increasing missing values from 0% to 60% using mean imputation (blue), hot-deck imputation (olive), MICE (purple) and EM (red) for categorical data. (A) describes the accumulated average bias if missing values are generated in a MAR fashion, (B) in a MCAR fashion. (C) shows the average bias if missing values are generated using the MAR mechanism and (D) the MCAR mechanism. Note that the error margin in (C) and (D) corresponds to the standard deviation of the samples estimates and not the bias. (Abbreviations: EM = expectation maximization, MAR = missing-at-random, MCAR = missing-completely-at-random, MICE = multiple imputation by chained equations)

RMSE performed significantly better than the other imputation methods ($p < 0.05$).

Figs 3 and 4 show the absolute bias. For numerical data MICE and EM showed the lowest absolute bias in the MAR case and mean and hot-deck imputation in the MCAR case. Mean imputation showed the lowest bias for categorical data for both MAR and MCAR. The mean imputation yielded less stable results for categorical data compared to numerical data.

Predictive model analysis

Fig 5 shows the performance of the GLM for an increasing amount of missing values that were generated in a MAR+MCAR fashion. Generally, the more values were imputed, the lower the resulting performance. The best overall performance was yielded by mean imputation (Fig 5A). Compared to the other imputation methods, this difference is significant only in the range of 25% to 45% missing values (Fig 5C). Listwise deletion showed the lowest performance compared to all other imputation methods (Fig 5B). Starting from 2% missing values every other imputation method performed significantly better than listwise deletion (Fig 5D).

The completed-entry model (0% missing values) showed higher AUC values compared to all imputation methods. The difference started to be significant for the

Fig 5. GLM performance on the dataset with increasing percentages of MAR+MCAR missing values and comparison of imputation methods with listwise deletion and mean imputation. (A) and (B) show the predictive model performance in terms of AUC for increasing MAR+MCAR missing values from range 0% to 60% and 0% to 10% respectively using mean (blue), hot-deck imputation (olive), MICE (purple), EM imputation (red) and listwise deletion (green). The plots in the bottom (C) and (D) show the corresponding *p* values of the different imputation methods compare to mean imputation (C) and listwise deletion (D) using a paired Wilcoxon signed-rank test. The horizontal dashed black line indicates 0.05, the threshold of significance for the *p* values. (Abbreviations: AUC = area under the curve, GLM = generalized linear model, EM = expectation maximization, MAR = missing-at-random, MICE = multiple imputation by chained equations)

MAR case-type between 2% to 3% missing values. From 11% on every model is significantly worse than the complete-entry model.

Similar results could be observed for MCAR missing values (Fig 6). The more values imputed, the lower the resulted AUC is. Mean imputation yielded the best performance, yet significance was shown only for 45% missing values and above (Figs 6A and 6C). Listwise deletion performed significantly lower than all other imputation methods starting from 3% missing values (Fig 6D).

Fig 6. GLM performance on the dataset with increasing percentages of MCAR missing values and comparison of imputation methods with listwise deletion and mean imputation. (A) and (B) show the predictive model performance in terms of AUC for increasing MCAR missing values from range 0% to 60% and 0% to 10% respectively using mean (blue), hot-deck imputation (olive), MICE (purple), EM imputation (red) and listwise deletion (green). The plots in the bottom (C) and (D) show the corresponding *p* values of the different imputation methods compare to mean imputation (C) and listwise deletion (D) using a paired Wilcoxon signed-rank test. The horizontal dashed black line indicates 0.05, the threshold of significance for the *p* values. (Abbreviations: AUC = area under the curve, GLM = generalized linear model, EM = expectation maximization, MCAR = missing-completely-at-random, MICE = multiple imputation by chained equations)

For the MCAR case-type, the completed-entry model performed the best as well. The first significant AUC value was for 1% missing values. Starting from 18% every model was significantly worse than the complete-entry model.

Discussion

In the present study, we developed a publicly available framework to investigate different imputation methods for handling missing values and tested it in a clinical stroke dataset as a use case. The utilized dataset 1000Plus represents a typical dataset in stroke regarding size and recorded values. For the predictive model, the results show that listwise deletion performs significantly worse than imputation methods starting from a low percentage (2% for MAR and 3% for MCAR). Additionally, our results indicate that for this type of data you should not impute data above 10% (MAR+MCAR) and 17% (MCAR). For the error assessment no method outperformed all other methods for every analysis. Furthermore, it seems to be crucial which missing value mechanism is underlying in the dataset.

Listwise deletion is still commonly practiced yet highly discouraged [1, 10]. Our results corroborate this notion and strongly suggest to use imputation methods. The performance of our predictive model started to drop significantly already when only 2% of the values were missing using listwise deletion. This implies that the available incomplete patient information still adds crucial value to the predictive model and should not be neglected.

Our results do not provide a strict recommendation for one imputation method. While mean imputation seemed to show the lowest RMSE and highest performance in terms of AUC, these results should be interpreted with caution. Mean imputation is a method that aims to reduce the RMSE, thus this measurement is biased towards mean imputation. Therefore, we additionally compared the methodologies using the absolute bias. Here, mean imputation performs well for categorical data as well as numerical data with MCAR missing values. Looking at the error assessment for categorical data, however, we observed that mean imputation performed less robustly. In the particular case of categorical data, mean imputation means imputing the value that occurs most often in the remaining dataset. Hence, the imputation method highly depends on which category the missing value belonged to. The resulting error is then less stable and more easily corrupted by the missing value pattern.

In the predictive model analysis, mean imputation showed significantly better results than other imputation methods in the range of 25% to 45% (MAR+MCAR) and 45% to 60% (MCAR) missing values. For the given dataset we establish a threshold of 11% (MAR+MCAR) and 18% (MCAR) over which imputation of missing values is discouraged. Consequently, the significant improvement of mean imputation is a priori not within the practical range where values should be imputed [26, 27].

For numerical data in the MAR case-type, we found MICE and EM to show the lowest absolute bias. In other studies, complex algorithms like MICE and EM also appeared to be superior to seemingly old-fashioned imputation methods like mean or hot-deck imputation [16, 17, 26]. In the case of numerical data and MCAR missing values, however, mean and hot-deck imputation showed the lowest bias. It seems unintuitive that simple algorithms like mean and sampling as the best performing imputation methods. The higher bias for MCAR compared to simpler imputation methods might, however, be explained by inherent characteristics of the more sophisticated methods. The MICE algorithm builds upon strong dependencies between the covariates. The missing value is estimated based on the corresponding values of the other parameters. When removing values in a completely random fashion, i.e. MCAR, the dependencies between the covariates might not be as strong anymore in our dataset. Thus, it is hard to reconstruct. If values are missing in a non-completely random fashion, i.e. MAR, there is pattern for missingness available that complex algorithms like MICE can learn from. The same holds true for the EM. The EM algorithm estimates the underlying log likelihood of the complete dataset [28, 29]. Given this distribution, the missing values are approximated. In the MAR case unlike MCAR, the existing pattern of missing values might help to capture the likelihood to yield a good estimation. To conclude, the underlying missing value mechanism might be very crucial regarding which imputation method is the most suitable for the given dataset.

In the error assessment, we observed that the error rate was quite constant for an increasing percentage of missing data. While appearing counterintuitive on first sight, the explanation for this phenomenon is simple: For each imputation method, we assess a value from a distribution that estimates the true underlying distribution. When drawing an increasing number of samples from both distributions, the difference between the values drawn from the two distributions remains the same on average. Thus, the error rate does not increase for more missing samples. Nevertheless, when looking at the accumulated error, we can see that for each imputed sample the new error is added so that the error rate is in fact increasing.

In the predictive model performance assessment, however, we see a different behavior. With increasing amount of missing values we can see decreasing predictive performance [30–32]. Depending on the size of the dataset and the number of covariates, the performance drops at a certain threshold. Thus, the significant decrease in performance can occur at a different percentage of missing values. In our study with 380 patients and 10 predictive parameters, the significant performance drop was measured starting from 11% missing values for MAR+MCAR missing values and 18% for MCAR respectively. Therefore, we suggest to only use imputation methods until 10% missing values. Our results confirm findings from the literature identifying numbers in a similar range of missing value percentages leading to a performance drop [26, 27].

Importantly, our results implicate that there is no generally "best" imputation method. Our findings suggest that – under certain circumstances – simple mean imputation might be superior to the other sophisticated imputation techniques. In other cases, i.e. MAR as the underlying missing value mechanism, MICE or EM performed better. This is corroborated also in theory by the ”no free lunch theorem” [33, 34]. The theorem states that there is no algorithm that performs best in all tasks. The good performance of one algorithm in one task comes with the cost of low performance in another task [33]. Since the imputation methods are in fact algorithms and the different dataset can be seen as tasks, the theorem could apply here as well. Hence, our results are specific for our dataset. Distinct characteristics of any other given dataset like its size, mechanism of data missingness and the type of features will influence which imputation method should be preferred. Thus, we make our framework publicly available (`https://github.com/tabeak/missing-value-analysis`). It can easily be used and adapted by other researchers to test their own datasets and identify the optimal imputation method for their data. Especially given the often limited size of datasets in medical applications, such an approach might allow increasing the validity of statistical testing and predictive modeling. Finally, our work is strongly encouraging further research on the performance of imputation methods in other tabular datasets.

Our study has several limitations. First of all, we could simulate MAR missing values only up to 9% due to mathematical constraints on covariates dependencies and the limited size of our dataset. Hence, our analysis mostly relates to mixed MAR+MCAR and MCAR mechanisms. The real underlying mechanism for missing values in clinical stroke datasets remains unknown. It is likely, however, that the true missing data mechanism is a mixture of MAR and MCAR as missing values can occur systematically as well as randomly in medical datasets. Secondly, due to data availability, we trained our predictive model on discharge mRS and not on final three months mRS, which is the clinically more useful measure. However, given the methodological nature of our study, the predictive model is only exemplary to show the impact of different methods dealing with missing data. The impact of missing data and different imputation methods on models predicting three months mRS must be elucidated in future studies.

Conclusion

We developed a publicly available R framework to evaluate different imputation methods and tested it on a typical clinical stroke dataset as a use case. Our main finding was that listwise deletion should not be performed and the choice of imputation methods might depend highly on the underlying missing value mechanism and other characteristics of a given dataset. Thus, we suggest that the optimal imputation method is dataset-dependent and we strongly encourage other researchers to adapt our openly available framework to their own datasets prior to analysis.

References

1. Wood AM, White IR, Hillsdon M, Carpenter J. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. International Journal of Epidemiology. 2005;34(1):89–99. doi:10.1093/ije/dyh297.

2. Burnett SJ, Deelchand V, Franklin BD, Moorthy K, Vincent C. Missing clinical information in NHS hospital outpatient clinics: prevalence, causes and effects on patient care. BMC health services research. 2011;11:114. doi:10.1186/1472-6963-11-114.

3. Rabideau DJ, Nierenberg AA, Sylvia LG, Friedman ES, Bowden CL, Thase ME, et al. A novel application of the Intent to Attend assessment to reduce bias due to missing data in a randomized controlled clinical trial. Clinical trials (London, England). 2014;11(4):494–502. doi:10.1177/1740774514531096.

4. Asadi H, Dowling R, Yan B, Mitchell P. Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy. PLoS ONE. 2014;9(2):e88225. doi:10.1371/journal.pone.0088225.
5. Livne M, Kossen T, Madai VI, Zaro-Weber O, Moeller-Hartmann W, Mouridsen K, et al. Multiparametric Model for Penumbral Flow Prediction in Acute Stroke. Stroke. 2017;48(7):1849–1854. doi:10.1161/STROKEAHA.117.016631.
6. Livne M, Boldsen JK, Mikkelsen IK, Fiebach JB, Sobesky J, Mouridsen K. Boosted Tree Model Reforms Multimodal Magnetic Resonance Imaging Infarct Prediction in Acute Stroke. Stroke. 2018;49(4):912–918. doi:10.1161/STROKEAHA.117.019440.
7. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. The New England journal of medicine. 2016;375(13):1216–1219. doi:10.1056/NEJMp1606181.
8. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion. 2019;50:71–91. doi:10.1016/j.inffus.2018.09.012.
9. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics. 2008;77(2):81–97. doi:10.1016/j.ijmedinf.2006.11.006.
10. Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Statistics in Medicine. 2003;22(4):545–557. doi:10.1002/sim.1340.
11. Liu Y, Gopalakrishnan V. An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data. Data. 2017;2(1). doi:10.3390/data2010008.
12. Higgins JP, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. Clinical Trials (London, England). 2008;5(3):225–239. doi:10.1177/1740774508091600.

13. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–525. doi:10.1093/bioinformatics/17.6.520.
14. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work? International journal of methods in psychiatric research. 2011;20(1):40–49. doi:10.1002/mpr.329.
15. King G, Honaker J, Joseph A, Scheve K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. American Political Science Review. 2001;95(1):21.
16. Barzi F. Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies. American Journal of Epidemiology. 2004;160(1):34–45. doi:10.1093/aje/kwh175.
17. Barzi F, Woodward M, Marfisi RM, Tognoni G, Marchioli R, Investigators oboGP. Analysis of the Benefits of a Mediterranean Diet in the GISSI-Prevenzione Study: A Case Study in Imputation of Missing Values from Repeated Measurements. European Journal of Epidemiology. 2006;21(1):15–24. doi:10.1007/s10654-005-5086-5.
18. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338(jun29 1):b2393–b2393. doi:10.1136/bmj.b2393.
19. Young-Saver DF, Gornbein J, Starkman S, Saver JL. Handling of Missing Outcome Data in Acute Stroke Trials: Advantages of Multiple Imputation Using Baseline and Postbaseline Variables. Journal of Stroke and Cerebrovascular Diseases: The Official Journal of National Stroke Association. 2018;27(12):3662–3669. doi:10.1016/j.jstrokecerebrovasdis.2018.08.040.
20. Hotter B, Pittl S, Ebinger M, Oepen G, Jegzentis K, Kudo K, et al. Prospective study on the mismatch concept in acute stroke patients within the first 24 h after symptom onset - 1000Plus study. BMC Neurology. 2009;9(1):60. doi:10.1186/1471-2377-9-60.

21. De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. BMC Medical Research Methodology. 2017;17(1). doi:10.1186/s12874-017-0372-y.

22. Eilaghi A, d'Esterre CD, Lee TY, Jakubovic R, Brooks J, Liu RTK, et al. Toward patient-tailored perfusion thresholds for prediction of stroke outcome. AJNR American journal of neuroradiology. 2014;35(3):472–477. doi:10.3174/ajnr.A3740.

23. Hand PJ, Wardlaw JM, Rivers CS, Armitage PA, Bastin ME, Lindley RI, et al. MR diffusion-weighted imaging and outcome prediction after ischemic stroke. Neurology. 2006;66(8):1159–1163. doi:10.1212/01.wnl.0000202524.43850.81.

24. van Seeters T, Biessels GJ, van der Schaaf IC, Dankbaar JW, Horsch AD, Luitse MJ, et al. Prediction of outcome in patients with suspected acute ischaemic stroke with CT perfusion and CT angiography: the Dutch acute stroke trial (DUST) study protocol. BMC Neurology. 2014;14(1):37. doi:10.1186/1471-2377-14-37.

25. Hiraga A, Yamaoka T, Sakai Y, Osakabe Y, Suzuki A, Hirose N. Relationship between outcome in acute stroke patients and multiple stroke related scores obtained after onset of stroke. Journal of Physical Therapy Science. 2018;30(10):1310–1314. doi:10.1589/jpts.30.1310.

26. Fellinghauer CS, Prodinger B, Tennant A. The Impact of Missing Values and Single Imputation upon Rasch Analysis Outcomes: A Simulation Study. Journal of Applied Measurement. 2018;19(1):1–25.

27. Acuña E, Rodriguez C. The Treatment of Missing Values and its Effect on Classifier Accuracy. In: Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation. Springer, Berlin, Heidelberg; 2004. p. 639–647. Available from: `https://link.springer.com/chapter/10.1007/978-3-642-17103-1_60`.

28. Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus. 2013;2(1):222. doi:10.1186/2193-1801-2-222.

29. Moon TK. The expectation-maximization algorithm. IEEE Signal Processing Magazine. 1996;13(6):47–60. doi:10.1109/79.543975.

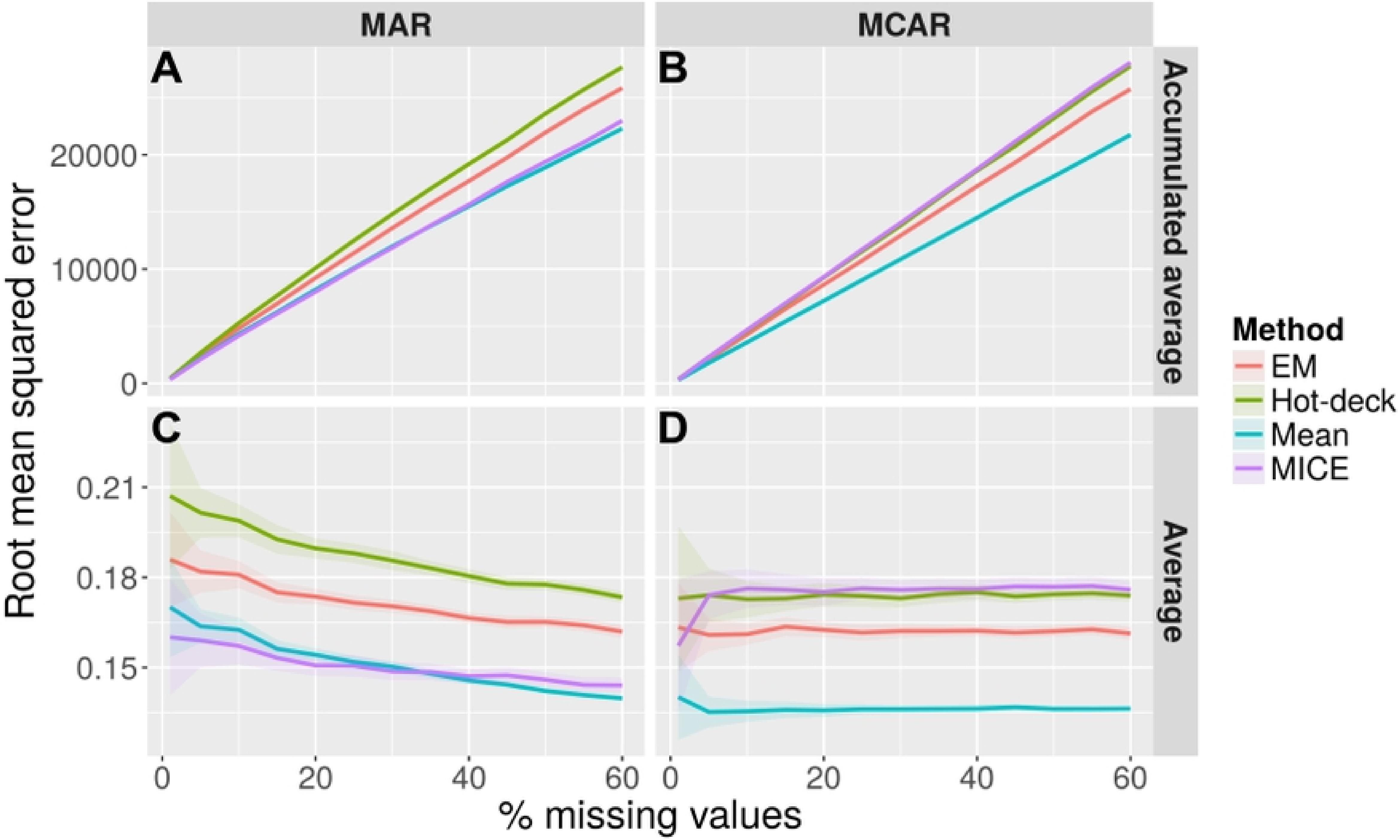
30. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making. 2012;12(1). doi:10.1186/1472-6947-12-8.

31. Kim SY. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. BMC bioinformatics. 2009;10:147. doi:10.1186/1471-2105-10-147.

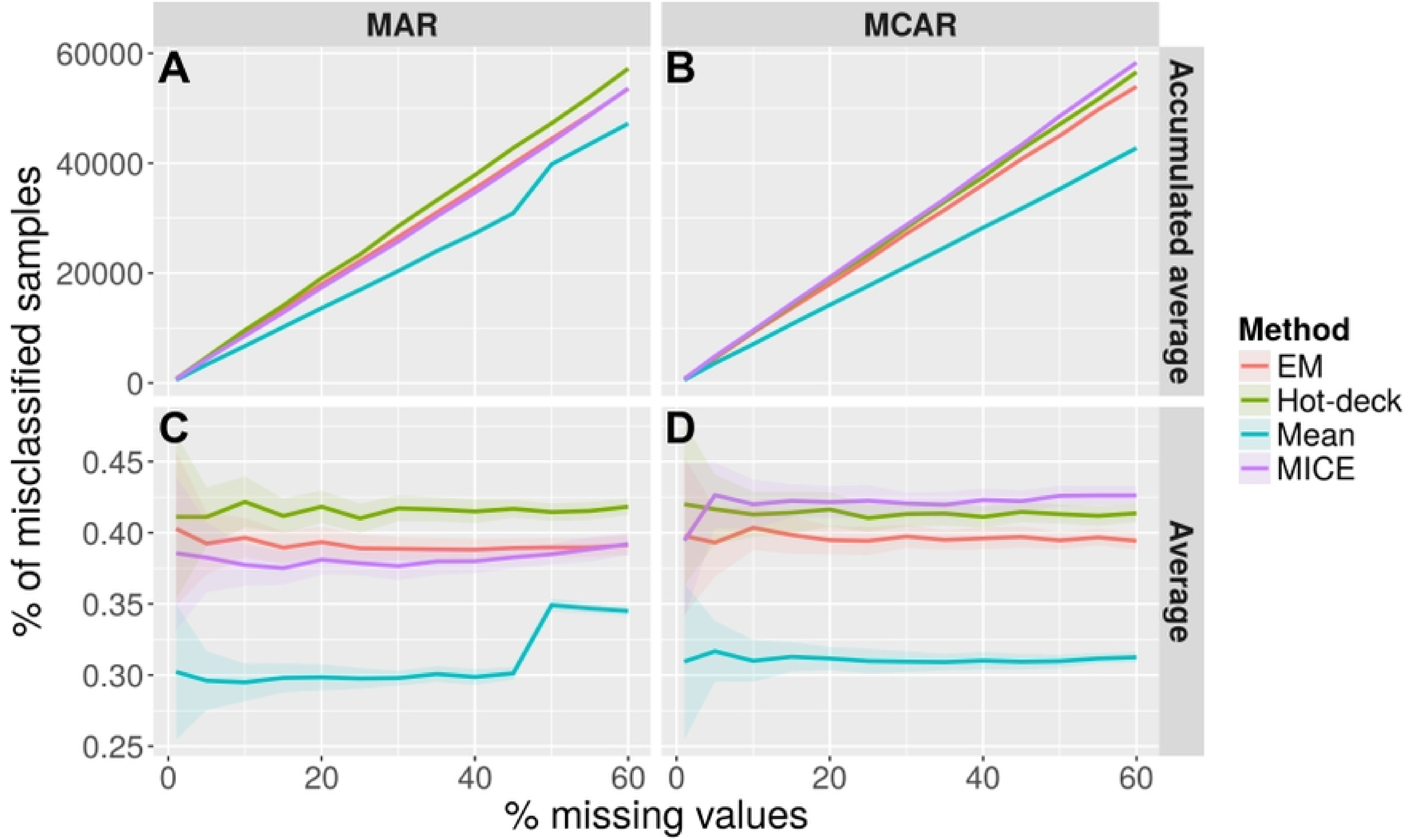
32. Kalayeh HM, Landgrebe DA. Predicting the Required Number of Training Samples. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1983;PAMI-5(6):664–667. doi:10.1109/TPAMI.1983.4767459.

33. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation. 1997;1(1):67–82. doi:10.1109/4235.585893.

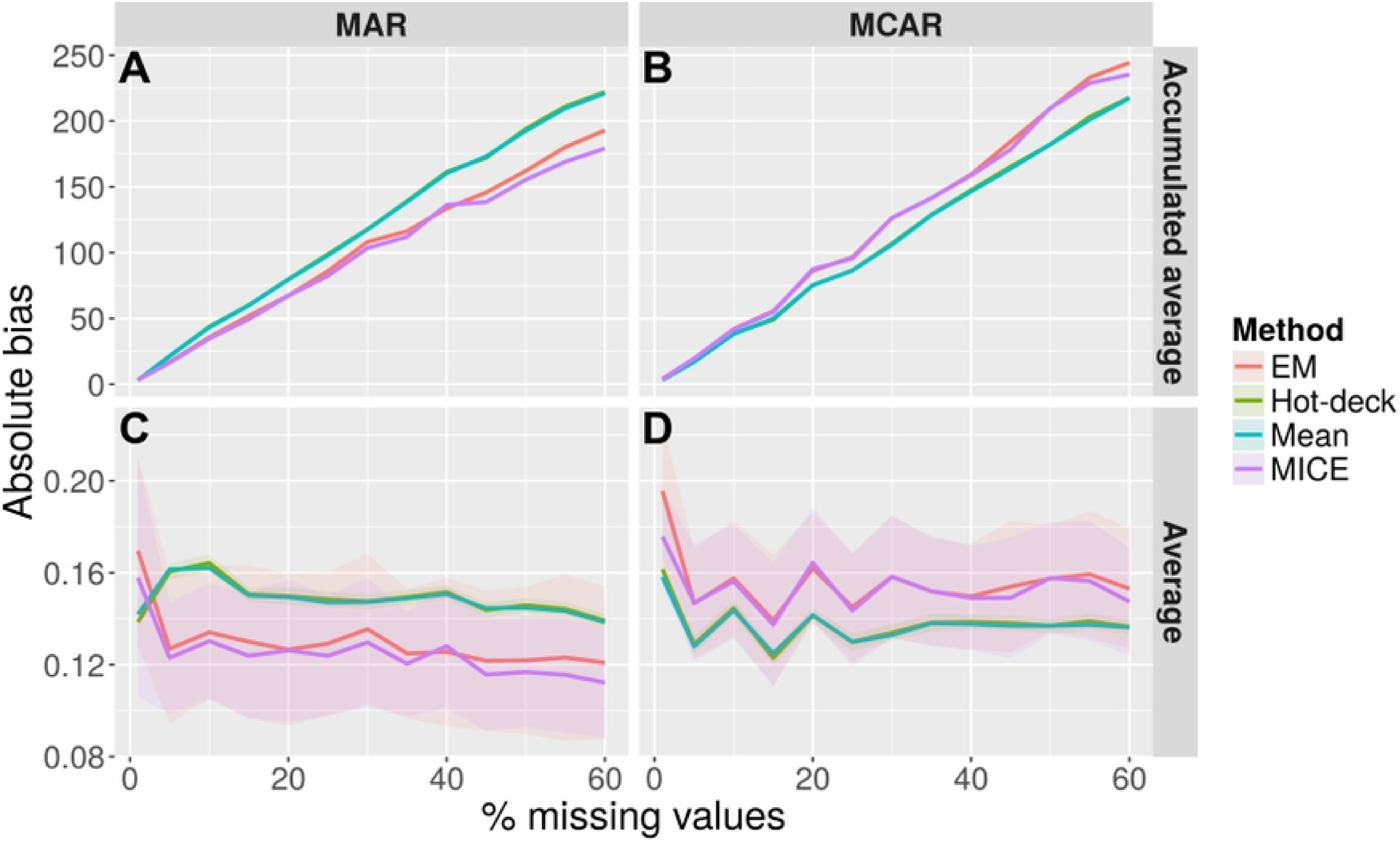
34. Wolpert DH, Macready WG. No Free Lunch Theorems for Search. Santa Fe Institute; 1995. Available from: `https://econpapers.repec.org/paper/wopsafiwp/95-02-010.htm`.


MAR
MCAR
A
B
C
D
Accumulated average
Average
Root mean squared error
20000
10000
0
0.21
0.18
0.15
0
20
40
60
% missing values
Method
EM
Hot-deck
Mean
MICE

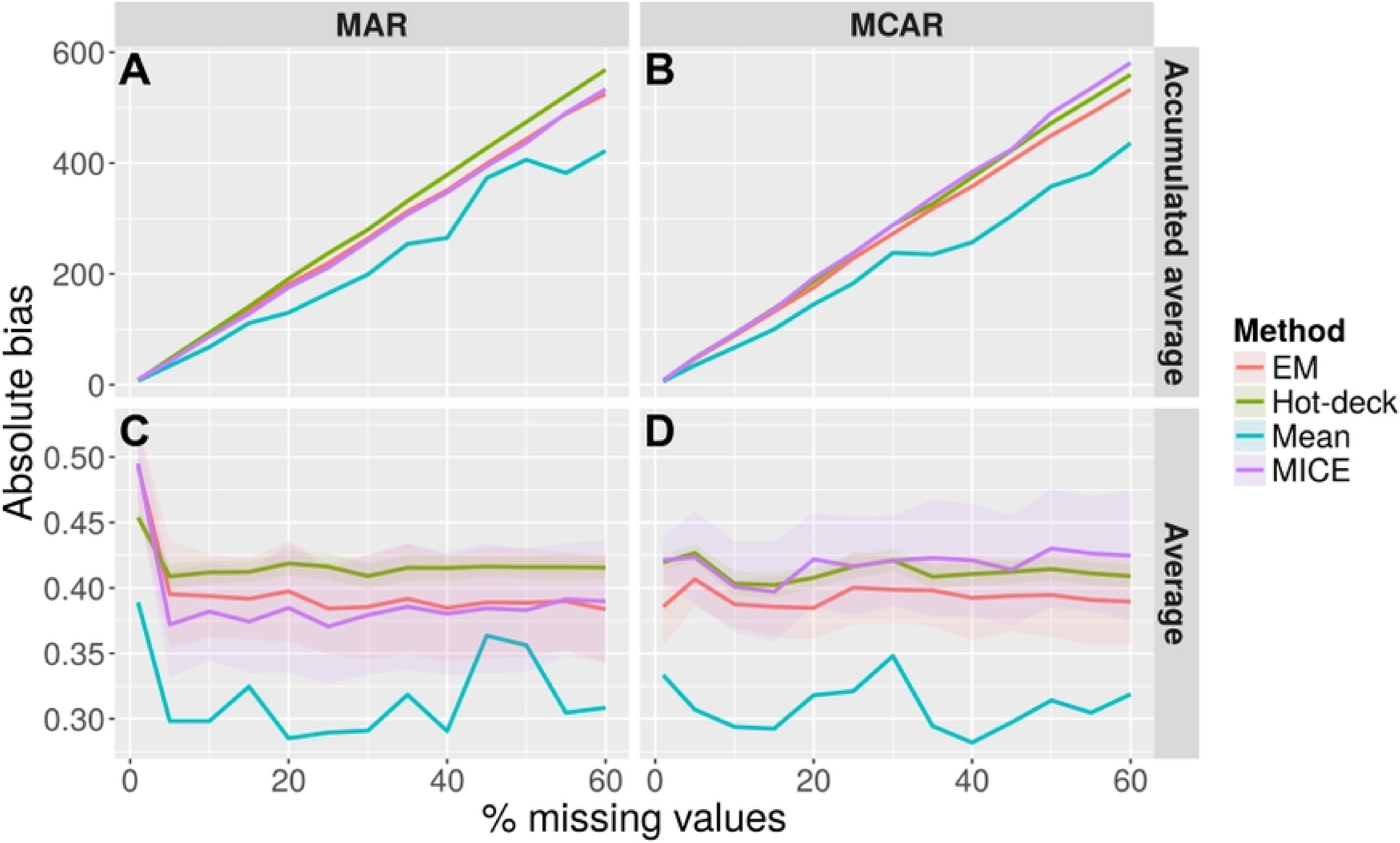
Figure


MAR
MCAR
A
B
C
D
Accumulated average
Average
% of misclassified samples
% missing values
60000
40000
20000
0
0.45
0.40
0.35
0.30
0.25
0
20
40
60
Method
EM
Hot-deck
Mean
MICE

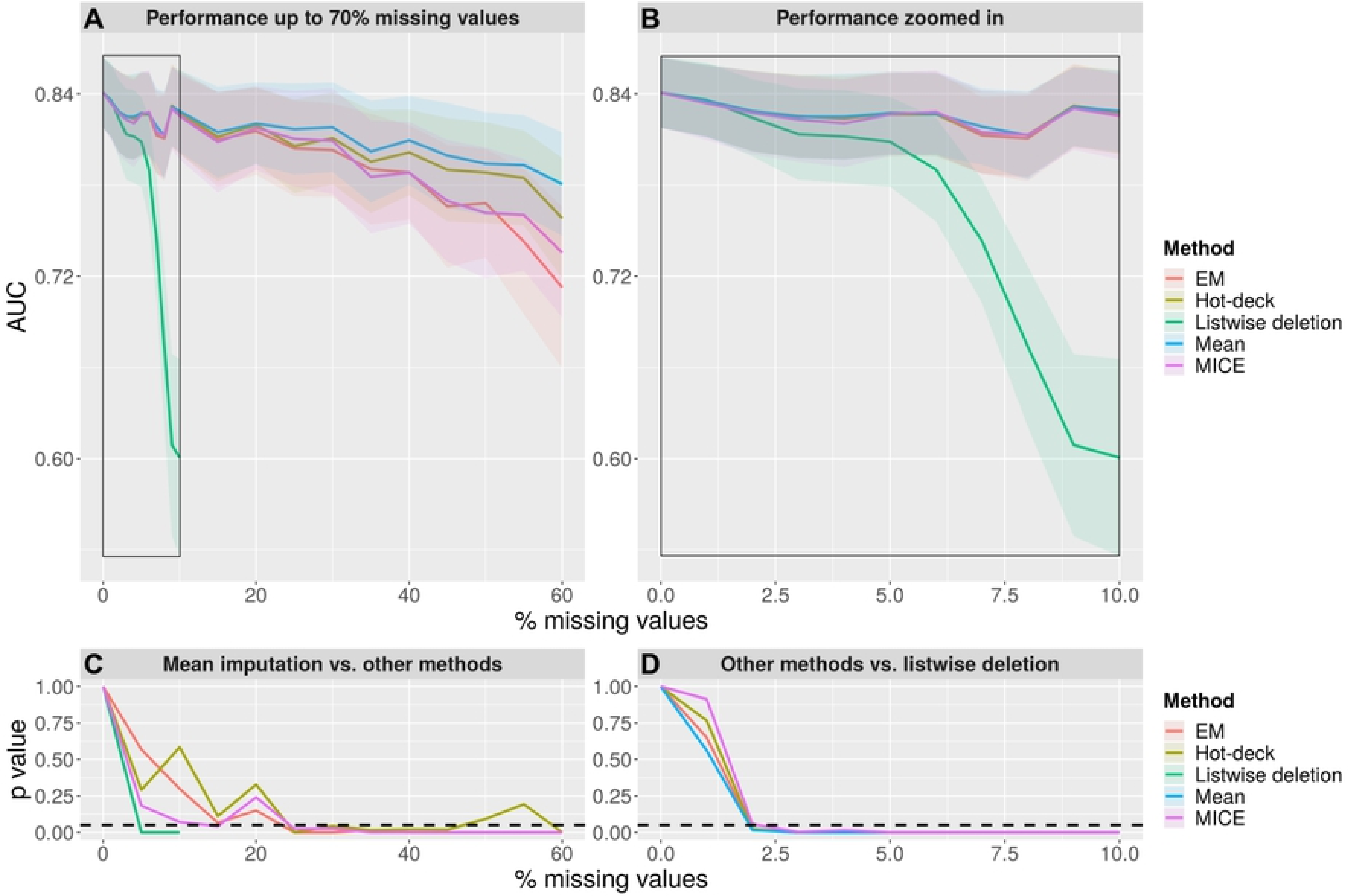
Figure


MAR
MCAR
A
B
C
D
Accumulated average
Average
Absolute bias
% missing values
0
50
100
150
200
250
0.08
0.12
0.16
0.20
0
20
40
60
Method
EM
Hot-deck
Mean
MICE

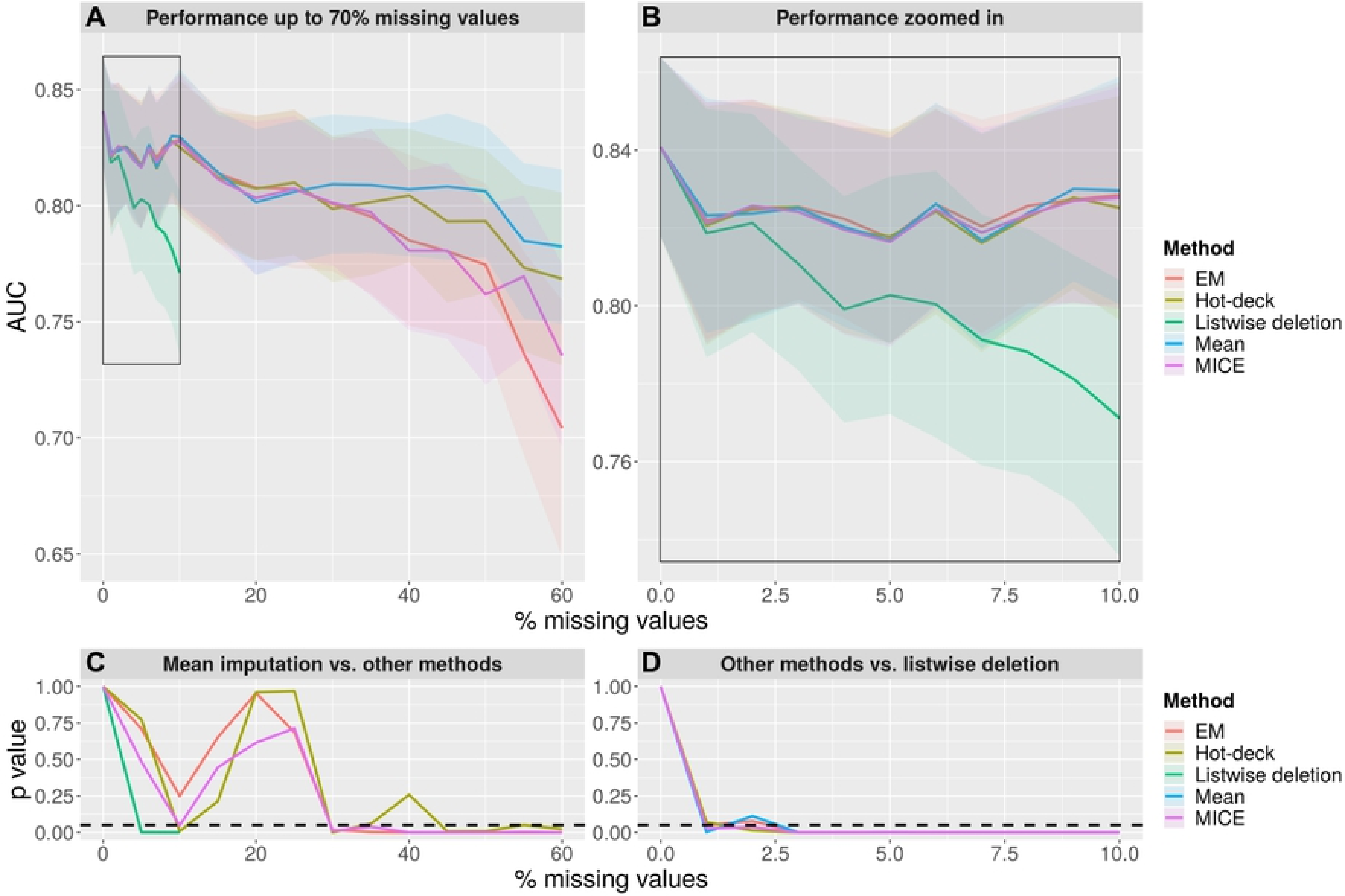
Figure


MAR
MCAR
A
B
C
D
Accumulated average
Average
Absolute bias
% missing values
600
400
200
0
0.50
0.45
0.40
0.35
0.30
0
20
40
60
Method
EM
Hot-deck
Mean
MICE

Figure


A
Performance up to 70% missing values
B
Performance zoomed in
AUC
0.84
0.72
0.60
0
20
40
60
0.0
2.5
5.0
7.5
10.0
% missing values
Method
EM
Hot-deck
Listwise deletion
Mean
MICE
C
Mean imputation vs. other methods
D
Other methods vs. listwise deletion
p value
1.00
0.75
0.50
0.25
0.00
% missing values
Method
EM
Hot-deck
Listwise deletion
Mean
MICE

Figure


A
Performance up to 70% missing values
B
Performance zoomed in
AUC
0.85
0.80
0.75
0.70
0.65
0.84
0.80
0.76
0
20
40
60
0.0
2.5
5.0
7.5
10.0
% missing values
Method
EM
Hot-deck
Listwise deletion
Mean
MICE
C
Mean imputation vs. other methods
D
Other methods vs. listwise deletion
p value
1.00
0.75
0.50
0.25
0.00
% missing values
Method
EM
Hot-deck
Listwise deletion
Mean
MICE

Figure