Can the Best Prediction Be the Worst Imputation?

On Common Measures to Evaluate Imputation Methods

Universiteit Utrecht

Boy Remmelzwaal

July 2021

First supervisor:
Dr. Gerko Vink

Second supervisor:
Prof. Dr. Stef van Buuren

MSc. Applied Data Science

Abstract

In the occurrence of missing data, handling the missingness through imputation can be a holy grail that allows for analyses of hypothetically complete data. The selection of an appropriate imputation method is important to prevent bias and wrong inferences. A measure commonly used to evaluate imputation method performance is to measure the difference between the imputed data and the true data, referred to as the root mean squared error (RMSE). However, the usefulness of using the RMSE may be questioned. The current study compares various imputation methods on commonly used performance measures to examine the differences between given performance indications. Data comes from simulations (nsim = 500) based on a case study. It is found that selecting an imputation method based on best performance according to RMSE would result in the selection of a method that leads to systematically biased regression coefficients, overestimating the proportion of explained variance and below nominal coverage rates. These results hold for various missingness scenarios. Recommendations are given on methods to evaluate imputation method performance and future use of the RMSE. The best predictions do not result in the best imputations.

Acknowledgements

I would like to take this brief moment to express my gratitude towards the people that have supported the journey of writing this thesis. Gerko, thank you kindly for your guidance and insightful stories that have sparked my interest in the topic. I can only admire your passion and knowledge about missing data and your willingness to share it! Hanne, thank you for your amazing feedback and helpful comments. Your enthusiasm and positive criticism reminded me to stay critical and really brought my work to a higher level.

I will follow the sun,
Boy Remmelzwaal.

Contents

1. Introduction

The problem of missing data is ubiquitous to almost any field of research. Yet, the topic of handling missing data is not often discussed outside of statistical journals (Altman & Bland, 2007). Missing values are argued to be especially common in medical and social science research (De Goeij et al., 2013; Rubin, 1996). To deal with missingness, many researchers use ad-hoc solutions such as complete case analysis, available-case analysis, or single-value imputations (Lang & Little, 2018; Pigott, 2001). However, unless researchers carefully consider the assumptions required for those methods, statistical inferences they make are likely to be biased and misleading (van Buuren 2018; Little & Rubin, 1987; Pigott, 2001). In a study assessing the handling of missing data in randomized clinical trials, Bell et al. (2014) found that 95% of the 77 identified studies reported some missing data. In most of these studies, missing data was handled using complete case analysis (45%), followed by single imputation (27%), and only 8% reported using multiple imputation (Bell et al., 2014). This indicates that despite the potential negative consequences, usage of such ad-hoc methods in practice remains most frequent. As studies continue to use ad-hoc methods, it is important to bring awareness to appropriate strategies to deal with missing data as well as highlight their effects on downstream analyses.

Throughout the years, various imputation methods have been developed to replace missing values. As not every method may result in statistically valid inferences it is important to evaluate the quality of imputations. A widely-used measure to evaluate the quality of these imputation methods is to compare the imputed data to the real data, also referred to as the (root) mean squared error ((R)MSE; Salfran, Jordan, Spiess, 2016). An RMSE of zero would imply no difference between the observed and imputed values. This suggests one did a perfect job at imputing the data. However, using the MSE may be considered too simplistic of a measure to evaluate how well the imputation is done (van Buuren, 2018). It is argued that this method "ignores the uncertainty of the missing values, resulting in biased estimates and invalid statistical inferences" (van Buuren, 2018, p. 56). Accordingly, what could be considered the best value prediction based on RMSE, may in reality be the worst imputation in terms of validity. For appropriate imputation methods and statistically valid inferences, it is therefore important to challenge this measure. Thereupon different imputation methods and their performance under common performance measures will be discussed.

Single imputation methods are those that "estimate what each missing value could have been and impute it with a single value" (Li, Stuart & Allison, 2015, p. 1966). Methods such as regression or mean imputation can be used as quick ad-hoc solutions to impute missing data, but can result in biased outcomes (van Buuren, 2018). Using mean imputation methods

will likely result in underestimation of the variance, affect the relations between variables and introduce bias to estimates other than the mean (van Buuren, 2018; Little & Rubin, 2002). Similarly, modeling the imputations using regression can result in biased correlations and underestimation of variability (van Buuren, 2018). Because single imputation does not account for the uncertainty of missingness, the standard errors computed from the imputed data are systematically underestimated, p-values are too small and confidence intervals too narrow (Little & Rubin, 2002).

Multiple imputation allows to introduce the uncertainty of the missing data by creating several imputed datasets, running statistical analyses on each, and pooling the results (Little and Rubin, 1987). The imputations are drawn from a posterior predictive distribution of the missing values (Rubin, 1996). Doing this captures the variability in the missing data, and therefore potentially generates more accurate estimates of what could have been (Vink, 2016). Multiple imputation is not a holy grail, as its validity depends on its assumptions. But in a context where those assumptions are met, Rubin (1996) firmly believes multiple imputation is the method of choice for addressing missing data and van Buuren (2018) describes many scenarios for which this is indeed the case.

As incorrect imputations have the potential to drastically affect downstream analyses and produce incorrect results (Graham, 2009), it is important to critically evaluate the performance of imputation methods. A few studies provided guidelines to evaluate imputation methods (e.g. Salfrán et al., 2016; Vink, 2016), but there is no consensus on which measures to use. Besides the RMSE, van Buuren (2018) proposes three other evaluation measures to assess the extent to which imputation methods obtain statistically valid inferences: raw bias, coverage rate (CR), and average width (AW). However, despite van Buuren's critical comments and demonstration on the deceptive nature of the RMSE, it is still a commonly used measure in evaluation studies (e.g. Junninen et al., 2004; Le, Beuran & Tan, 2018; Zakaria & Noor, 2018). As such it can be argued that additional work is required that points out guidelines for imputation method evaluation.

Similar to van Buuren's (2018) concerns, prior research has addressed that some performance measures of imputation methods may be misleading (Salfran, Jordan, Spies, 2016). Salfran et al. (2016) demonstrated this by pointing out that properties of the downstream analysis based on IRMI (a robust imputation function) are systematically worse compared to those based on multiple imputations by chained equations (MICE). These results were in contrast to the mean values of the error measures; in all but one case the error measure was smaller for the imputations based on IRMI. Consequently, if one would take an imputation method that

leads to biased estimators and coverage rates, this could result in falsely rejecting a true null hypothesis far too often (Salfran et al., 2016). Extensive amounts of studies have discussed the comparison of methods to deal with missing data. But in comparison, methodological implications and evaluation of performance measures is a gap few have addressed.

The current paper aims to compare various imputation methods for handling missing data and evaluate their validity with a variety of performance measures. To evaluate the performance of these techniques, the raw bias, coverage rate, reverse model bias, and root mean squared error will be used. This allows us to reflect on the effect of using different performance measures on deciding what is the most appropriate imputation method. Besides, different levels of missingness will be considered and evaluated to see whether these findings hold for varying missingness scenarios. Model-based simulation with parameters derived from real-world data will be used to compare the true data to imputed values. Results of the study may be used to provide quantitative evidence that what appear to be statistically good imputation methods according to some measures, are in reality not as valid. Accordingly, this paper will answer the following research questions:

1. Which imputation methods perform best given common performance measures?
2. What is the effect of different levels of missingness on the imputation methods proposed performance measures results?

2. Methods

2.1 Data

To address the questions, a simulation study was conducted. The data was simulated based on the PIMA Indians dataset (UCI Machine Learning, 2016). This dataset contains eight diagnostic measurements relating to diabetes. Variables were selected based on their correlations to retain some relational structure in the data ($r > .1$). Specifically, four predictor variables were selected: Glucose, BloodPressure, SkinThickness, Insulin, and target variable BMI. This resulted in the following model of analysis:

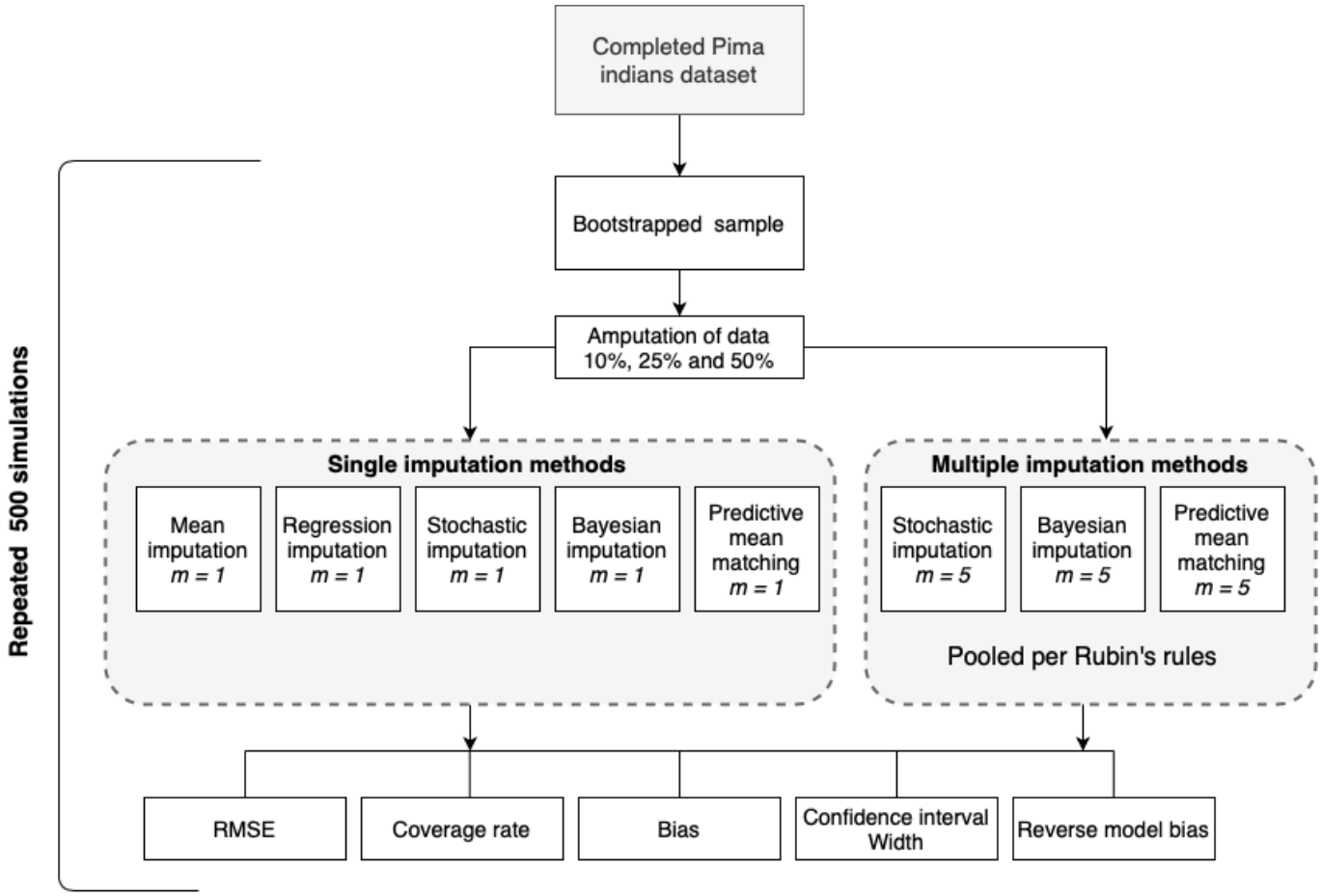
$$\hat{BMI} = \alpha + \hat{\beta}_1 Glucose + \hat{\beta}_2 BloodPressure + \hat{\beta}_3 SkinThickness + \hat{\beta}_4 Insulin + \hat{\epsilon}$$

Though metadata indicated no missing data, researchers have pointed out biologically implausible zero values, i.e. blood pressure values of zero (Pearson, 2006). As a solution for the current context, zero values were treated as NA and imputed to generate a single complete data set, which was used as the ground truth. All data simulations and analyses were conducted using R (R Core Team, 2021). For the imputations, the **mice** package (van Buuren & Groothuis-Oudshoorn, 2011) was used.

2.2 Simulation

A total of 500 simulations were conducted (nsim = 500). For every simulation, the complete data was bootstrapped to mimic the sampling process (n = 792). Next, the bootstrapped dataset was made incomplete by amputing the dataset using four multivariate missingness patterns (Schouten, Lugtig & Vink, 2018). The data was made missing according to missing completely at random missingness (MCAR). This implies that the probability to be missing was equal for all cases. In all four missingness patterns, a maximum of 2 variables were made missing. In total three scenarios were used with the proportion of missing cases being 50%, 25%, and 10%. These proportions of missingness are suggested to at least be considered when evaluating imputation routines (Vink, 2016). The data was imputed using the methods described in Table 1. An overview of the simulation can be found in Figure 1.

Figure 1: Flowchart of the simulation study.



2.3 Description of imputation methods

Missing data was imputed using eight different methods (see Table 1). For the multiple imputation methods, five imputations were used (m = 5). This is argued to be sufficient under moderate levels of missingness, and can therefore be considered a reasonable standard value (van Buuren, 2018). For the multiple imputation methods, seven iterations were used which, based on previous findings, is deemed sufficient (Raghunathan, Solenberger & van Hoewyk, 2002; Oberman, van Buuren & Vink, 2020). An overview and descriptions of all used imputation methods can be found in table 1.

Table 1: Overview of imputation methods.

Imputation method	imputations m	Description
Mean	1	Replacing missing values with the mean value of a variable, calculated from all other respondents.
Regression	1	Other independent variables are used to produce a regression equation of the missing variable. Missing values are replaced with predictions from the regression equation.
Stochastic linear regression	1, 5	Similar to regression imputation, but introduces noise to the imputation by adding a residual error term.
Bayesian linear regression	1, 5	Similar to stochastic regression, but introduces parameter uncertainty by drawing the parameters from their posterior distributions given the data (van Buuren, 2018).
Predictive mean matching	1, 5	Missing values are replaced with values drawn from respondents whose regression predicted score is close to the regression score of the respondent (Landerman, Land & Pieper, 1997). Matching with d = 5 is the default in MICE (van Buuren, 2018).

2.4 Analyses and performance measures

Following the statistical measures discussed by van Buuren (2018), the performance of the imputation methods was analyzed using the raw bias, RMSE, the coverage rate (CR), and the confidence interval width (CIW). In addition, the reverse model bias was evaluated. Results were averaged across all simulations. Following, descriptions of the measures are given.

Raw bias. First, the raw bias of an estimate can be defined as the difference between the expected value of the estimate $\bar{Q}$ and its true value Q (van Buuren, 2018). The closer the value to zero, the better.

$$Raw\ Bias = E(\bar{Q}) - Q$$

RMSE. Next, the RMSE may be calculated by taking the squared error of the imputed value minus the actual value. Doing this evaluates the estimate of a parameter Q on accuracy as well as precision (van Buuren, 2018).

$$RMSE = \sqrt{(E(\bar{Q}) - Q)^2}$$

Coverage Rate and Confidence Interval Width. I calculated bootstrap confidence intervals for all regression coefficients using the bootstrap standard errors, i.e the square root of the variance of the vector of estimates. This results in the empirical standard deviation of the realized bootstrap sampling distribution of the estimates and more appropriate confidence intervals. The coverage rate (CR) indicates the proportion of simulations in which the confidence interval of an estimate holds the true value. A confidence interval of 95% is used for which the closer to a nominal coverage of 95%, the better (Vink, 2016). The confidence interval width (CIW) is the distance between the upper bound and the lower bound of the 95% confidence interval. CR naturally increases as the width of the confidence interval gets larger, except for scenarios with larger bias. A narrower confidence interval implies better statistical efficiency. Depending on the context, a tradeoff must be made between coverage rate and confidence interval width to determine what is an appropriate imputation method.

Reverse model bias. To assess whether the imputation methods captured the characteristics of the original dataset, the model was "flipped" by regressing the outcome on each of the predictors. Accordingly, reversing the model resulted in four new models having each of the original predictors as outcome variable. Valid imputation should be able to result in unbiased regression coefficients for any model in the data.

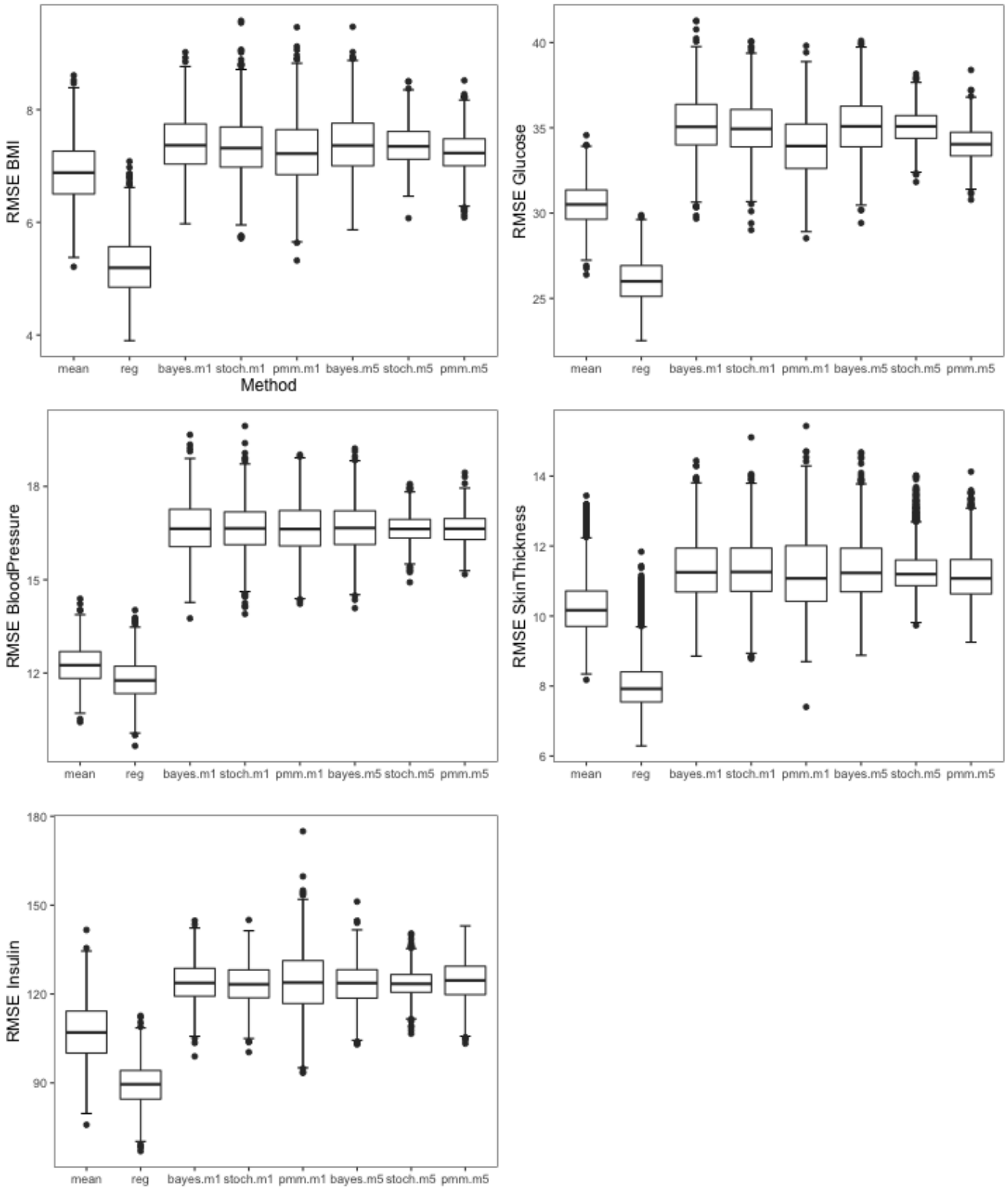
3. Results

The following section presents the results for each imputation method over 500 simulations. First, I present the RMSE results. Following, I present the proportion of explained variance, bias, coverage rate, and confidence interval width to evaluate the performance of imputation methods for the 50% missingness scenario. Note that for each imputation method the regression coefficients and performance measures are calculated by fitting the model of analysis shared in the methods section. Next, I share the results of the reversed model fits. Finally, results of the 10% and 25% missingness scenarios are considered.

3.1 RMSE

Figure 2 visualises the RMSE for the variable estimates across all simulations. Uniformly for all variables, regression imputation results in the lowest RMSE. Mean imputation has the second-best performance according to the RMSE. Other methods have comparable results for RMSE across all variables. Based on the RMSE, regression imputation would be preferred as the difference between the predicted value and true value is the smallest.

Figure 2: RMSE of imputations for 50% missingness.



note. Mean = mean imputation, regr. = regression imputation, bayesm1 = bayesian single imputation, stochm1 = stochastic single imputation, pmmm1 = predictive mean matching single imputation, bayesm5 = bayesian multiple imputation (M = 5), stochm5 = stochastic multiple imputation (M = 5), pmmm5 = predictive mean matching multiple imputation (M = 5).

3.2 Bias

Table 2 displays the average bias of the regression estimates compared to the true estimates. The most severely biased estimates for all predictors are found when using mean imputation. Regression imputation resulted in the second-worst biases. This implies that mean and regression imputation had the greatest tendency to over-or underestimate the regression coefficients.

Methods other than mean and regression imputation produced relatively similar biases and were overall found to be less biased. The best performance varied per coefficient. The least biased methods per regression coefficient have been made bold in Table 2.

Table 2: Average simulation estimates, bias, and R-squared for 50% missingness.

	Glucose	BloodPr.	SkintTh.	Insulin	
		Coefficient			R^2
True values	-0.00567	0.08593	0.37966	0.00823	0.4597
		Bias			
mean	0.0065	-0.0087	-0.0536	0.0006	0.3476
regr.	0.0047	0.0073	0.0288	0.0003	0.5154
stoch.m1	-0.0005	-0.0013	0.0022	**0.0000**	0.4664
bayes.m1	0.0002	-0.0020	0.0008	**0.0000**	0.4654
pmm.m1	0.0009	**-0.0005**	0.0010	-0.0001	0.4635
stoch.m5	**0.0001**	-0.0018	0.0021	**0.0000**	0.4646
bayes.m5	-0.0003	-0.0020	0.0015	**0.0000**	0.4641
pmm.m5	0.0009	**-0.0005**	**0.0005**	-0.0002	0.4640

note. Mean = mean imputation, regr. = regression imputation, bayesm1 = bayesian single imputation, stochm1 = stochastic single imputation, pmmm1 = predictive mean matching single imputation, bayesm5 = bayesian multiple imputation (M = 5), stochm5 = stochastic multiple imputation (M = 5), pmmm5 = predictive mean matching multiple imputation (M = 5). Best bias values are bold.

3.3 Proportion explained variance

Table 2 also displays the variance explained by the coefficients in the model (R^2). It was found that mean imputation resulted in underestimation of the variance explained (R^2 = .35). This is because part of the predictive power is lost due to the missingness and not maintained with mean imputation. Contrasting, regression imputation results in overestimation of explained variance (R^2 = .52). As with regression imputation the imputations are calculated under the fitted model, the relations get strengthened. Other imputation methods had approximately unbiased results, with slight overestimation of the model variance.

3.4 Coverage rate and confidence interval width

Table 3 displays the standard errors, confidence interval widths, and coverage rates of the coefficient estimates using a 95% confidence interval. SEs and CIWs are only relevant to consider in case there is nominal coverage ($\geq$ 95%). Mean imputation yields systematic undercoverage across the estimates. Regression imputation resulted in undercoverage for the estimates of Glucose, SkinThickness, and Insulin, but has nominal coverage for BloodPressure. Other methods were found to have coverages close to nominal coverage for all estimates. Multiple imputation methods had coverages above nominal (¿ 95%) for all variables, implying the methods were slightly too conservative.

CIW's for multiple imputation methods were found to be slightly narrower compared to other methods. Stochastic multiple imputation had the smallest intervals given good coverage. Accordingly, mean and regression imputation are least reliable and stochastic multiple and single imputation seems very efficient.

Table 3: Average confidence interval coverage results for 50% missingness.

	Glucose			BloodPressure			Skinthickness			Insulin		
	S.E	CIW	Coverage	S.E	CIW	Coverage	S.E	CIW	Coverage	S.E	CIW	Coverage
mean	0.0080	0.0313	0.872	0.0241	0.0944	0.936	0.0300	0.1177	0.554	0.0024	0.0093	0.956
regr.	0.0131	0.0513	0.928	0.0329	0.1292	0.956	0.0358	0.1406	0.874	0.0032	0.0126	0.934
stoch.m1	0.0121	0.0475	**0.950**	0.0293	0.1151	0.962	0.0344	0.1352	**0.956**	0.0032	0.0126	0.952
bayes.m1	0.0125	0.0492	0.938	0.0290	0.1141	0.940	0.0346	0.1358	0.958	0.0032	0.0127	**0.950**
pmm.m1	0.0130	0.0512	0.958	0.0292	0.1145	**0.954**	0.0340	**0.1335**	0.962	0.0033	0.0128	0.962
stoch.m5	0.0109	**0.0451**	0.972	0.0275	**0.1131**	0.968	0.0335	0.1341	0.966	0.0029	**0.0119**	0.962
bayes.m5	0.0109	0.0457	0.972	0.0277	0.1154	0.968	0.0335	0.1347	0.966	0.0029	**0.0119**	0.970
pmm.m5	0.0110	0.0463	0.964	0.0278	0.1156	0.962	0.0339	0.1386	0.968	0.0029	0.0121	0.960

note. Mean = mean imputation, regr. = regression imputation, bayesm1 = bayesian single imputation, stochm1 = stochastic single imputation, pmmm1 = predictive mean matching single imputation, bayesm5 = bayesian multiple imputation (M = 5), stochm5 = stochastic multiple imputation (M = 5), pmmm5 = predictive mean matching multiple imputation (M = 5). Smallest interval given good coverage are bold. Best coverages are bold.

3.5 Reverse model bias

Table 4 presents the bias of the coefficient estimates in the reversed models. Mean imputation resulted in systematic bias, having the most severe biases for BMI's effect on Glucose, BloodPressure, and SkinThickness. Regression imputation resulted in the most biased estimates for the effect of BMI on Insulin as was the second most biased method in the other models.

Regarding the best results, the performances of the methods varied per model. Stochastic multiple imputation was found to be least biased for the effect of BMI on Glucose as well as on SkinThickness. Next, predictive mean matching single imputation was found to be least biased for the effect of BMI on BloodPressure. Finally, bayesian multiple imputation was found

to be least biased for the effect of BMI on Insulin. Accordingly, mean and regression imputation were most biased in the reverse model, whereas stochastic and bayesian multiple imputation were least biased.

Table 4: Reversed model regression coefficient biases for 50% missingness.

	BMI →Glucose	BMI →BloodPr.	BMI →SkinTh.	BMI →Insulin
		Coefficient		
True values	-0.1287	0.4548	0.9731	2.5841
	bias	bias	bias	bias
mean	0.1458	-0.1585	-0.149	-0.1452
reg	-0.0676	-0.0154	0.0419	-0.2702
stoch.m1	-0.0106	-0.0066	0.0048	-0.0246
stoch.m5	**0.0001**	-0.011	0.002	-0.0323
bayes.m1	0.0065	-0.0114	**0.0004**	-0.032
bayes.m5	-0.0047	-0.0117	0.0011	**-0.0175**
pmm.m1	0.0226	**-0.0058**	-0.0033	-0.0271
pmm.m5	0.0204	-0.0068	-0.0031	-0.0274

note. Mean = mean imputation, regr. = regression imputation, bayesm1 = bayesian single imputation, stochm1 = stochastic single imputation, pmmm1 = predictive mean matching single imputation, bayesm5 = bayesian multiple imputation (M = 5), stochm5 = stochastic multiple imputation (M = 5), pmmm5 = predictive mean matching multiple imputation (M = 5). Best bias values are bold.

3.6 Different levels of missingness

In addition to 50% proportion of incomplete cases, scenarios with 10% and 25% were evaluated to see if they were sensitive to differences. As 50% missingness has been priorly discussed, the focus will be on 10% and 25% missingness only. For reasons of brevity only the results for variable Glucose are given. Glucose was selected as moderate differences between methods had been found under the 50% scenario. Results of other variables can be found on GitHub[1].

Table 5 shows the estimates and performance measures for every scenario. Summarising, most results concerning RMSE, bias, CR and CIW are in concordance with the 50% missingness scenario. A deviating result in the 10% and 25% scenarios was that bias gradually increased for mean and regression imputation as the missingness increased. This in contrast to all other methods, which were not affected by greater missingness. Besides that, under 10% missingness, there were no notable differences between confidence interval widths. As the missingness percentage increases, the CIW of mean imputation decreases. This means there is less variation in the estimate. Contrasting, other imputation methods increase in CIW with no notable difference between respective methods. This could be explained by the variance

[1]https://github.com/boy-r/Thesis

of imputations being larger than the actual values. Accordingly, greater missingness seems to decrease the efficiency of imputations.

All in all, regression imputation would be the best method according to RMSE for all missingness scenarios. Yet, based on bias, mean and regression imputation would be worst biased in all missingness scenarios. Confidence interval coverage is close to nominal for every method, though regression imputation had the lowest coverage for the 10% scenario and mean imputation for the 25% and 50% scenarios.

Table 5: RMSE, bias, and confidence interval coverage results for 10%, 25%, and 50% missingness scenarios.

			10% missingness			
	estimate	S.E	bias	CIW	Coverage	RMSE
mean	-0.0040	0.0088	0.0016	0.0347	0.944	30.3
regr.	-0.0068	0.0095	-0.0011	0.0374	0.936	**22.62**
stoch.m1	-0.0055	0.0094	0.0002	0.0368	0.944	31.66
bayes.m1	-0.0055	0.0093	0.0002	0.0365	0.942	31.72
pmm.m1	-0.0053	0.0096	0.0004	0.0375	**0.950**	30.6
stoch.m5	-0.0056	0.0091	**0.0001**	0.0357	0.948	31.82
bayes.m5	-0.0056	0.0092	**0.0001**	0.0362	0.940	31.89
pmm.m5	-0.0053	0.0092	0.0003	0.0362	0.942	30.94
			25% missingness			
	estimate	S.E	bias	CIW	Coverage	RMSE
mean	-0.0025	0.0085	0.0032	0.0332	0.928	30.48
regr.	-0.0085	0.0104	-0.0029	0.0409	0.944	**22.78**
stoch.m1	-0.0059	0.0100	-0.0003	0.0392	0.952	31.62
bayes.m1	-0.0057	0.0101	**0.0000**	0.0397	0.952	31.72
pmm.m1	-0.0053	0.0101	0.0004	0.0398	**0.950**	30.98
stoch.m5	-0.0058	0.0094	-0.0001	0.0376	0.942	31.72
bayes.m5	-0.0057	0.0094	**0.0000**	0.0378	0.946	31.79
pmm.m5	-0.0052	0.0097	0.0005	0.0390	0.944	30.97
			50% missingness			
	estimate	S.E	bias	CIW	Coverage	RMSE
mean	0.0008	0.0080	0.0065	0.0313	0.872	30.43
regr.	-0.0104	0.0131	0.0047	0.0513	0.928	**23.24**
stoch.m1	-0.0062	0.0121	-0.0005	0.0475	**0.950**	31.68
bayes.m1	-0.0055	0.0125	0.0002	0.0492	0.938	31.87
pmm.m1	-0.0047	0.0130	0.0009	0.0512	0.958	30.95
stoch.m5	-0.0057	0.0109	**0.0001**	0.0451	0.972	31.78
bayes.m5	-0.0059	0.0109	-0.0003	0.0457	0.972	31.92
pmm.m5	-0.0048	0.0110	0.0009	0.0463	0.964	31.15

note. Mean = mean imputation, regr. = regression imputation, bayesm1 = bayesian single imputation, stochm1 = stochastic single imputation, pmmm1 = predictive mean matching single imputation, bayesm5 = bayesian multiple imputation (M = 5), stochm5 = stochastic multiple imputation (M = 5), pmmm5 = predictive mean matching multiple imputation (M = 5). Best values under measure are bold.

4. Discussion

The current study found that different performance measures lead to widely different conclusions about the performance of imputation methods. Comparing the RMSE values, it was found that using regression imputation resulted in the lowest RMSE, and thus would be the best method under this measure. However, based on the other performance measures different conclusions were made: the estimates were not valid. Regression and mean imputation were found to result in the most severely biased regression coefficients. Flipping the models resulted in similar results, with mean and regression imputation having the largest bias. This implies that those imputation methods were not able to maintain the characteristics of the original dataset. Moreover, mean imputation and regression imputation also resulted in the most bias in R^2 (Variance explained). Finally, mean and regression imputation were found to be least confidence valid, as they had the lowest coverage rates across regression coefficients. In contrast, other methods were found to have greater RMSE values, but were overall less biased whilst having valid confidence intervals. Therefore, it has been shown that using the RMSE as an imputation performance measure can give a wrong perception of how good an imputation method actually is.

In addition, it has been examined whether the difference between these performance measures changes under different proportions of missingness. Results were no different compared to 50% missingness. RMSE values were found lowest for regression imputation under both 10% and 25% missingness, but mean and regression imputation were also found to be most biased in both scenarios. This implies that selecting an imputation method based on RMSE would result in biased coefficient estimates regardless of low or high percentages of missingness. Though there was no clear best performing imputation method across scenarios, in the current context stochastic or baysian multiple imputation would yield more statistically valid imputations.

One could question why the method that is best at recreating the true data, is not the best method after all. And why the RMSE, which measures the discrepancy between the true and imputed data, would not be informative of good imputation. The main problem is that imputing the best value under model fit (i.e. regression imputation) does not take into account the uncertainty of missing values. The results showed that imputing the best value under model fit results in substantially different regression weights compared to real data. This caused variance to be underestimated, correlations to be biased upwards and the proportion of explained variance to be overestimated. Underestimating the variance means it becomes more likely one will find spurious results. Accordingly, p-values will be too low resulting in researchers and practitioners to draw conclusions based on relationships that are artificially inflated.

These findings are in line with the results of Salfran et al. (2016). In their paper,

selection of imputation methods based on the normalised root mean squared error resulted in preference for a method that led to biased estimators and severely downward biased coverage rates. Furthermore, the findings support the argument made by van Buuren (2018) that the RMSE is not informative for evaluating the quality of imputation methods. The current study adds to prior findings by providing an additional case that challenges the RMSE as a measure for imputation evaluation. Besides, the proposed reverse model bias measure used in this study is a relatively unused but helpful indicator that could be a great addition for future literature.

Limitations and future research

A potential limitation of this study is the scope, which is limited to MCAR missingness. Though MCAR is the minimum consideration for imputation methods (Vink, 2016), the missingness mechanism can affect the imputation quality and produce different results. However, as mean and regression imputation usually perform worse under missingness at random (MAR; Scheffer, 2002), it is expected that the current findings hold true for MAR contexts. Nonetheless, future research could focus on extending the findings to MAR missingness to confirm this. Another limitation is that the results of the study pertain to the specific case study used. Based on confirmation of previous studies (van Buuren, 2018; Salfran et al., 2016), it is assumed that findings extend to other data as well. Finally, future imputation method evaluation studies could benefit from additional literature which can be used as as a northern star in terms of which measures to use.

Recommendations for practice

The RMSE is one of the most widely used performance indicators in imputation research(Schmitt, Mandel & Guedj, 2015). I propose that researchers must be cautious with the use of RMSE or similar measures (i.e root-mean-square Deviation and mean absolute error), as it can be a misleading indicator of imputation method performance. Current and prior research showed that method selection merely based on RMSE may favor methods that are biased in downstream estimates and increase the rate of false positives (van Buuren, 2018; Salfran et al., 2016). Instead, based on the current study I recommended to select methods based on coverage rates, confidence interval width and raw bias as proposed by van Buuren (2018). In addition I recommend using the reverse model bias as an extra measure of imputation validity. Depending on the context of research, one has to make a tradeoff whether there is more value in being less biased, or having better confidence interval width and accompanying coverages. Finally, I recommend not to use regression and mean imputation to avoid biased results. The best prediction can, indeed, be the worst imputation.

5. References

Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C. H. (2014). Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology, 14(1), 1-8.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R.Journal of statistical software, 45(3), 1-68. DOI: 10.18637/jss.v045.i03

van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

de Goeij, M. C., van Diepen, M., Jager, K. J., Tripepi, G., Zoccali, C., & Dekker, F. W. (2013). Multiple imputation: dealing with missing data. Nephrology Dialysis Transplantation, 28(10), 2415-2420.

Vink, G. (2016). Towards a standardized evaluation of multiple imputation routines.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38(18), 2895-2907.

Landerman, L. R., Land, K. C., & Pieper, C. F. (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods & Research, 26(1), 3-33.

Le, T. D., Beuran, R., & Tan, Y. (2018, November). Comparison of the most influential missing data imputation algorithms for healthcare. In 2018 10th International Conference on Knowledge and Systems Engineering (KSE) (pp. 247-251). IEEE.

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. Jama, 314(18), 1966-1967.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data (No. 519.5 L778). J. Wiley.

Little, R. J., & Rubin, D. B. (2002). Single imputation methods. Statistical analysis with missing data, 59-74.

Oberman, H.I., van Buuren, S. & Vink, G. (2020). Missing the Point: Non-Convergence in Iterative Imputation Algorithms. First Workshop on the Art of Learning with Missing Values (Artemiss) hosted by the 37 th International Conference on Machine Learning (ICML).

Pearson, R. K. (2006). The problem of disguised missing data. Acm Sigkdd Explorations Newsletter, 8(1), 83-92.

Pigott, T. D. (2001). A review of methods for missing data. Educational research and evaluation, 7(4), 353-383.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raghunathan, T. E., Solenberger, P. W., Van Hoewyk, J. (2002). IVEware: Imputation and variance estimation software. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434), 473-489.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley Sons.

Salfrán, D., Jordan, P., Spiess, M. (2016). Missing data: On criteria to evaluate imputation methods. Tech. rep. Discussion Paper.

Scheffer, J. (2002). Dealing with missing data.

Schmitt, P., Mandel, J., Guedj, M. (2015). A comparison of six methods for missing data imputation. Journal of Biometrics Biostatistics, 6(1), 1.

Schouten, R. M., Lugtig, P., Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation, 88(15), 2909-2930.

UCI Machine Learning. (2016, October). Pima Idians Diabetes Database, Version 1. Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database

Zakaria, N. A., Noor, N. M. (2018). Imputation methods for filling missing data in urban air pollution data formalaysia. Urbanism. Arhitectura. Constructii, 9(2), 159.

Unused “captionsetup[1] on input line