Biometrical Journal 52 (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Towards a standardized evaluation of imputation methodology

Hanne I. Oberman*,[1] and Gerko Vink [1]

[1] Departement of Methodology & Statistics, Padualaan 14, 3584 CH Utrecht, The Netherlands

Received zzz, revised zzz, accepted zzz

Developing new imputation methodology has become a very active field. Unfortunately, there is no consensus on how to perform simulation studies to evaluate the properties of imputation methods. In this paper we propose a move towards a standardized evaluation of imputation methods. To demonstrate the need for standardization, we highlight a set of potential pitfalls that bring forth a chain of potential problems in the objective assessment of the performance of imputation routines. This may lead to suboptimal use of multiple imputation in practice. Additionally, we suggest a course of action for simulating and evaluating missing data problems.

Key words: Evaluation; Imputation; Missing data; Simulation studies;

Supporting Information for this article is available online from github.com/gerkovink/StandardizedEvaluation.

1 Introduction

Imputation is a state-of-the-art technique for drawing valid conclusions from incomplete data. The technique has earned a permanent spot in research and policy making, demonstrated e.g. by the detailed manual created by the National Research Council (Little et al., 2012). Although a top-down enforcement of valid ways to handle missing data is not yet very pronounced, an increasing amount of researchers are embracing imputation techniques. After all, the principle of imputation is very intuitive.

The idea behind imputation is to impute (fill in) missing data, to obtain a valid estimate of what could have been. A popular variant is multiple imputation (Rubin, 1976), whereby each missing value is imputed multiple times. The completed data that are thus obtained can be analyzed by standard techniques and, in the case of multiple imputation, the analysis results can be combined into a single inference (using Rubin's rules Rubin, 1987). In contrast to ad hoc methods for dealing with missing values (e.g. list-wise deletion, mean imputation, regression imputation, last observation carried forward, indicator method), multiple imputation properly takes into account the sources of uncertainty that are related to the missingness problem.

The quality of a solution obtained by imputation depends on the statistical properties of the incomplete data and the degree to which an imputation procedure is able to capture these properties when modeling missing values. In general it holds that modeling missing data becomes more challenging when the amount of missingness increases [ref?]. However, when (strong) relations in the data are present, the observed parts can hold great predictive power for the models that estimate the missingness. In that case, imputation would be substantially more efficient than the ubiquitous list-wise deletion.

When evaluating the statistical properties (and thereby the practical applicability) of imputation methodology, researchers most often make use of simulation studies. In such studies a complete dataset is usually generated from a statistical model, another model is used to induce missingness, and a set of evaluation criteria is postulated to evaluate the performance of one or

*Corresponding author: e-mail: h.i.oberman@uu.nl

more missing data methods. However, no golden standard has been established to evaluate imputation routines and, as a result, the validity of the simulation procedures may differ tremendously from one developer to another. Especially with novel imputation methods being propagated from the fields of machine learning and artificial intelligence, the differences may become more pronounced. Although these promising new methods seem to yield even sharper imputations than now-standard (semi-)parametric imputation methods, the comparison may not be fair due to the simulation setup.

The purpose of this paper is threefold: First, to raise some concerns with respect to evaluating imputation methodology. These concerns stem from careful consideration with fellow 'imputers' and from encounters as a reviewer for statistical journals. Second, to provide imputation methodologists with a suggested course of action when simulating missing data problems. This suggested approach should identify a common ground, but is in no way intended as an absolute solution. This identifies the third purpose of this paper: discussion. We hope to elicit critical thinking with respect to the problems at hand. We are all convinced that our methodology has some merit. But for sake of progress it would be much more advantageous if the aim of our evaluations would go beyond proving the point and would legitimately consider the statistical properties.

2 Why some evaluations should not be trusted

As of today, there is no consensus on how to perform simulation studies to evaluate the properties of imputation methods. Typically, the developer of a new imputation routine does some tests by simulations, but these tests differ across developers. This brings forth a chain of potential problems in the objective assessment of imputation method performance, within and across studies, which may lead to sub-optimal use of imputation in practice. To demonstrate the broad impact of these problems, we subdivide the problems in the following three distinct categories: problems with data generation, problems with missingness generation and problems with performance evaluation. We further detail the impact these problems may have on the validity of the performance evaluation of the imputation routines [this sentance could be smoother].

2.1 Data generation problems

To evaluate the ability of an imputation routine to handle missingness, a form of truth has to be established. Those who perform simulation studies are in the luxury position to establish the truth beforehand by choosing a data generating mechanism. Data generating mechanisms define how a complete dataset is obtained at the start of each simulation repetition. There are two general approaches to generaring complete data: (i) model-based simulation, in which data are drawn from a known statistical model or probability distribution, such as the multivariate normal distribution; and (ii) design-based simulation, where data are sampled (with replacement) from a sufficiently large observed set, such as from official statistics records.

The problem with model-based data generating mechanisms is that a method's performance on simulated data may not translate to empirical data. Real-life data hardly ever follow a given theoretical distribution, so there is no guarantee that simulation results are generalizable. Moreover, data are often generated such that the problem that is being studied is most pronounced, e.g. with consistently high correlations between groups of variables. This results in simulated data that contain such valuable information structures that, no matter what type of missingness is subsequently induced, the observed parts of the data still hold much (if not all) of the information about the missing part. Unsurprisingly, the performance of any imputation method will then be evaluated as good. Another threat to the generalizibility of model-based simulations is the use of a single model for both data generation and imputation. If data are generated following a model that is also used for imputing the data, the imputation approach will be deemed good (or better

than other methods) purely due to the evaluated conditions being in favor of the problem that is studied. Other (unfair) comparative advantages in favor of a certain imputation method may occur due to characteristics of the generated data, such as the number of observations, the number of variables, the variable type(s) and the coherence between variables. In contrast to design-based studies, such characteristics are not always explicit simulation conditions, which may give a false sense of objectivity.

An obvious problem with design-based simulation is that obtaining a large dataset without missing entries can be very challenging. Most real-world data contains at least one missing entry, for which the true underlying non-response model is by default unknown. Therefore, the simulator needs to deal with missingness in some way before incomplete empirical data can serve as comparative truth in the simulations. It may seem as an intuitive solution to only draw complete cases from the large dataset, which would indeed yield complete samples. However, there may be inherent differences between cases with and cases without any missing values, due to the unknown non-response model. Only sampling complete cases would thus result in a simulation that fails to capture all relevant real-world conditions [which refutes the main reason for using a design-based data generating mechanism in the first place]. Another way to deal with missingness in a design-based simulation is to impute the incomplete dataset once, to obtain a single completed dataset to draw samples from. Unfortunately, this practice may favor the imputation method that was used in this initial imputation step throughout any further evaluations. Nowadays, ... [Add that some simulation studies do not use a 'ground truth' at all, they just look for the method with the best predictive performance for a certain completely observed target variable. These studies apply different combinations of imputation and estimation methods on one or more benchmark datasets, and assess the RMSE of the predicted values after imputation and estimation to say something about the comparative performance of the methods. In this context, there is no talk of retrieving the 'true' but unobserved values at all. Refer to missing data chapter (Liu et al., 2021).]

Something else to consider is the number of samples that is drawn from the data generating mechanism. If evaulating missing data methodology is the primary focus of the simulation, it may not be necessary to draw data repeatedly. Sampling variation has always been an essential part of the evaluation of multiple imputation methodology. However, in order to obtain information about a method's ability to handle the missing data problem, or to objectively compare methods on their ability to correct for missingness, it is not necessary to take sampling variation into account (Vink and van Buuren, 2014). After all, we are interested only in the missing data mechanism, and are not considering the noise induced by the sampling mechanism for evaluation in such studies. Note that the conventional pooling rules (cf. Rubin, 1987, p. 76-77) do not apply for finite population inference. Instead, alternative pooling rules need to be used (Raghunathan et al., 2003; Vink and van Buuren, 2014). [An example from our own experience is that it may be difficult to get true parameter estimates of transformed data (in this case, the mean of a variable with a skew normal distribution). If we wouldn't have used a simulation set-up without sampling variance, it may have led to imprecise reference values of the estimands in our simulation. So, on the one hand, the typical set-up obfuscates the process that we're interested in (i.e., varying the missing data problem over simulation repetitions), while on the other hand, SEs may be incorrect if the sampling variance is not handled correctly in subsequent inferences.]

2.2 Missingness generation problems

In order to evaluate the effectivity of imputation methodology in recovering the established truth-from the data generating mechanism, some form of missingness has to be induced. Often, however, reports of simulation studies remain vague about the actual missingness conditions under investigation and, even worse, some authors only report something like

We generated missing data following a missing at random missingness mechanism.

This should be considered unacceptable as claims about the validity of the imputation inference heavily depend on the simulated missingness conditions, such as missingness patterns and missingness mechanisms. Missingness patterns concern the location of missing entries across variables in an incomplete dataset, whereas missingness mechanisms describe the relationship between missingness and the values of variables in the data (Little and Rubin, 2020, p. 8). Under this definition, the missingness pattern also encompasses the proportion of missingness in a given dataset.

Even the terminology on missingness generation can be confusing. Does a missingness proportion of 50% mean that half of the entries in an incomplete dataset are missing, or that half of the rows have at least one missing entry? In this paper, we will henceforth refer to the latter as the proportion of incomplete cases, and keep the term 'missingness proportion' restricted to the variable-by-variable interpretation. This distinction, however, is not always clear in the literature. Convoluting the two concepts may lead to incorrect generalizations from simulation studies, since the two are rarely equal (e.g., a proportion of 50% incomplete cases could translate to a missingness proportion of just 25% in bivariate data, or even 10% in data with five variables). Moreover, the value of the actual missingness proportion may be diffusing too. Some studies use only 10% missing data [rephrase to use missingness proportion?] where other studies push the limits to additionally investigate the performance under at least 50% missing data. This inconsistent display of simulation results may impact the objectivity of meta-evaluations over imputation methods, as one method's performance may appear to be favorable because of the less stringent simulation conditions. This ultimately may lead to statisticians recommending a less efficient method to applied researchers, thereby limiting the efficiency of the imputation approach and unnecessarily lowering the statistical power.

Other problems with missingness patterns entail the distribution of missing values across cases (as opposed to the proportion over variables). [Add: monotone/univariate vs multivariate, sporadic vs systematic, and univariately generated missing data may not have the expected multivatiate patterns. For example, iterative imputation algorithms converge instantly if there is only 1 variable with missingness or monotone pattern, but require iteration in all other situations.] Specific missing data patterns may inadvertently guide which imputation method prevails in performance. Some ad hoc methods are known to yield valid inferences under certain restricted conditions (e.g., list-wise deletion may outperform imputation methods if there is missingness in the outcome variable of the analysis model exclusively), while biasing inferences in any other case. Extrapolating simulation results from such a specific missing data pattern to more intricate empirical missingness patterns could lead to false conclusions.

[Missingness mechanisms: 1) definitions of mechanisms, 2) that MCAR is often ignored, while it's the minimum requirement and reference, and sometimes realistic in practice, 3) that MAR can be generated spuriously if the correlation between variables is low, and 4) that MNAR is often ignored, while it may be the most realistic in practice, 5) and finally something about types of M(N)AR (ref dance paper rianne).]

In order to obtain valid imputation inference, the imputation model must capture the essence of the true non-response mechanism (Meng, 1994). The model–if any–that is used to generate the missingness is usually assumed to be random (MAR) or completely random (MCAR).

[MCAR is important but often ignored:] MCAR is a necessary simulation condition for evaluating the performance of imputation procedures. Under MCAR, the statistical properties of the observed data given the missing data are known and any imputation routine that cannot at least mimic the performance of the observed data inference, should be deemed inefficient in the scope of the simulation. If an imputation method is not able to solve the problem (i.e. yield valid inference) under MCAR, the statistical properties of the procedure are not sound. Sadly, the straightforward case of MCAR is often neglected from simulation studies and focus is drawn to

Table 1: Missingness mechanisms.

Mech.	Interpretation and consequences of the mechanism
MCAR	The probability to be missing is the same for all cases. In other words, the missing values are missing at random and the observed values are observed at random.
MAR	The probability to be missing is the same within groups of cases defined by the observed data only. The essence is that observed relations in the data inform the missingness, such that MCAR may be assumed within the observed groups.
MNAR	The probability to be missing dependend on unobserved aspects of the data. The cause of the missingness is unknown and cannot be inferred from the observed data. This is considered non-ignorable missingness (see e.g. Rubin, 1976).

the evaluation of MAR mechanisms only. Alternatively, some studies limit their evaluations to MCAR mechanisms only, which in our view may be far too simplistic.

[MAR may be spurious:] With MAR missingness mechanisms, observed relations in the data are used to induce missingness during simulation (e.g. weight is made incomplete based on gender) such that the probability to be missing is the same within groups of cases defined by the observed data (e.g. males are less likely to disclose their weight, but gender is observed). The relations in the observed data may, however, be weak or non-existent. If MAR is induced based on weaker relations in the data, claims for a method's applicability to situations where the missingness is random become less valid. In the most extreme case, MAR is induced from data without multivariate relations. The inferential implications of the missingness would then mimic those of MCAR, even though the missingness is random. In fact, this would amount to one of the special cases under which complete case analysis would be more efficient than multiple imputation (see e.g. Van Buuren, 2012, p. 48): the missingness does not depend on the incomplete variable. This property might be useful in practice, but considering it as a condition to evaluate performance under MAR missingness is useless.

[MNAR is often ignored:] Missingness mechanisms that are not random (MNAR) and mechanisms that are considered non-ignorable (see e.g. Rubin, 1976), are generally ignored during evaluation of imputation methodologies, except for those methods that are specifically targeted at non-ignorable applications. Although one cannot definitively verify if the missingness is random–after all, for every MNAR mechanism there is a MAR mechanism with equal fit (Molenberghs et al., 2008)–it can be argued that MNAR is the more likely mechanism for real life missingness scenarios.

[Missingness type is often ignored (duplicated text from solutions section): Given the simulated data-distributions, one random missingness model may be far more disastrous to the observed information than another model. This may influence the performance of some (but not necessarily all) imputation routines. For example, inference from hot-deck techniques such as predictive mean matching (Little, 1988; Rubin, 1986) may be more severely impacted by large amounts of one-tailed missingness than inference from parametric techniques. It would be a shame to overlook such results due to the focus on a single MAR mechanism.

2.3 Evaluation problems

[Add bridge from last section. And explain how this part is subdivided: first problems with evaluation of imputations (general characteristics of the imputations such as convergence and plausibility), then evaluation of performance (in terms of actual performance measures).]

[All imputation models should be congenial, but this is difficult to test to often ignored even in single imputation endeavor, let alone in simulation studies with hundreds of imputations. Checks of imputation model fit are often omitted. Ignoring congeniality may yield sub-optimal imputations, which may disadvantage certain methods compared to others. Same goes for convergence: FCS is iterative and requires algorithmic convergence, but this is typically evaluated through visual inspection which is unfeasible in simulation studies. Problems with convergence may be even more severe: some methods do not produce imputations at all for certain conditions, while at other times the imputation algorithm is cut off too soon, and performance will be under-estimated.]

Moreover, the performance of imputation procedures on distributional properties is often ignored in simulation studies, and even though the estimates on the analysis level may be justified, some methods can yield imputations that may seem completely invalid to applied researchers. For example, one could very accurately estimate average human height by filling in negative values and values that are unrealistically large. While the obtained inference could still be valid under such imputations, the plausibility of the imputed values given the observed data should be under scrutiny.

[Add segway between evaluation of imputation and actual performance measures.] The evaluation criteria used to assess imputation performance vary from one developer to another. This is not surprising as people from different fields could have a different focus on the problem at hand. But the choice of performance measure(s) may inadvertently distort statistical properties of the imputed data.

[Add choice of analysis model/estimand here, since performance measure are defined wrt the estimand, e.g., focusing on univariate estimates may favor invalid ad hoc methods.]

Developers often only inquire about the 'accuracy' (i.e. how well can the method reproduce the original data). [Add RMSE at cell level here?] The goal of multiple imputation is not to reproduce the data, but to allow for obtaining valid inference given that the data are incomplete. We are interested in the correct answer to the research question; not in the truth itself. This means that, given the framework provided by Rubin (1987), statistical properties such as bias, confidence intervals and the coverage rate of the confidence intervals should be studied. After all, the 95% confidence interval should contain the 'true' value at least 95 out of 100 times (Neyman, 1934, p. 591). [So, what could go wrong by focusing on accuracy exclusively? See problems with RMSE of predictions.]

[Add problems with RMSE here. And that if the goal is inference (not prediction) then you need appropriate uncertainty, so m = 1 does not suffice.]

3 Suggested course of action

[In geveral, adhere to ADEMP (Morris et al., 2019). But there are some differences for the specific case of imputation methodology. All of the parts in ADEMP are used, but a bit scrambled up. Table 1 outlines some of the steps to consider in simulation studies focused on evaluating imputation methodology. The rest of this section is ordered based on these steps.]

3.1 Set scope

[Make aim clear. Choose scope (type of DGM, target) and imputation methods to evaluate. Ideally, perform fully factorial design, see Morris et al. (2019). Think about tolerable level of uncertainty to pick a number of simulation repetitions. Think about general set-up: are the missing data methods all applied to the same incomplete dataset in each simulation repetition, or is a new incomplete set created per method?]

Table 2: Steps to consider in imputation simulation studies.

Step	Aspects under consideration
1. Set scope	Aim(s), missing data method(s) to evaluate, number of simulation repetitions
2. Obtain truth	Data generating mechanism(s), estimand(s), sampling variance
3. Induce missingness	Missingness mechanism(s), missing data pattern(s), missingness proportion(s)
4. Apply methods	Imputation model(s), analysis model(s)
5. Evaluate imputations	Algorithmic convergence, distributional characteristics, imputation model fit
6. Evaluate performance	Performance measures,
7. Report	Text, visualization, checklist, simulation script

3.2 Obtain truth

[Decide which data genetation mechanism to use and obtain true value of the estimand(s).]

In model-based simulation data are drawn from a known probability distribution, so the theoretical parameters under which the samples are obtained serve as the comparative truth in the simulations. In design-based simulation, data are sampled from a sufficiently large set, so the parameters of this specific set serve as the comparative truth in the simulations.

The design-based approach is often used in situations where a probability distribution is not available, or where real-life data structures are of interest. Applications of design-based simulation are often found in official statistics. A benefit of design-based simulation is the ability to use real-life observed data structures.

For evaluations of model-based simulations, it could be convincing to demonstrate a method's applicability to real-world missing data problems. This can, for example, be achieved by obtaining and imputing a fully observed set of variables, wherein the missingness is mimicked from a similar, incomplete set. Alternatively, an incomplete set could be obtained and truth could be established by removing the incomplete cases from the data. However, the real-world missingness would then be omitted.

The simulator must to decide whether there is a need for sampling variance in the simulations scheme. For situations where sampling variance is not of interest, the sampling process can be omitted and a single complete dataset can simply be obtained from the data generating mechanism. The parameters of that single complete set will serve as comparative truth. This process is computationally convenient, because only a single complete datasets has to be considered during all of the simulations. Use alternative pooling rules for finite population inference (Raghunathan et al., 2003; Vink and van Buuren, 2014). If sampling variance cannot be omitted from the simulation scheme, draw data repeatedly and apply conventional pooling rules (cf. Rubin, 1987, p. 76-77).

3.3 Induce missingness

[Missingness mechanisms] First, one should always consider MCAR missingness, i.e. the scenario where the missing values are missing at random and the observed values are observed at random. Under MCAR, the statistical properties of the observed data given the missing data are known and any imputation routine that cannot at least mimic the performance of the observed data inference, should be deemed inefficient in the scope of the simulation.

Next, missing data should be induced conform a model that is dependent on the observed data. A straightforward technique for inducing different forms of univariate MAR missingness is described in Van Buuren (2012, p. 63) and a generalization to multivariate MAR missingness can be found in Buuren et al. (2006, Appendix B) and Brand (1999, §5.2.3). If the missingness is to be induced in longitudinal data, autoregressive MAR models (e.g. cf. Shara et al., 2015, model 2 and model 3) can be useful.

[Although often ignored, it can be argued that MNAR is the more likely mechanism for real life missingness scenarios. It would be wise to include the mechanism in your simulation study, or perform sensitivity analyses.] An indication of the validity of the obtained inference, given that the assumed missingness mechanism is suspected to be invalid, may be obtained by performing sensitivity analysis (see e.g. Molenberghs et al., 2014, part 5).

[Missingness type] Third, it is advisable to investigate varying shapes of MAR missingness to achieve a more realistic indication of the robustness of the imputation performance across the range of random missingness. Given the simulated data-distributions, one random missingness model may be far more disastrous to the observed information than another model. This may influence the performance of some (but not necessarily all) imputation routines. For example, inference from hot-deck techniques such as predictive mean matching (Little, 1988; Rubin, 1986) may be more severely impacted by large amounts of one-tailed missingness than inference from parametric techniques. It would be a shame to overlook such results due to the focus on a single MAR mechanism.

[Missingness proportion] Fourth, the amount of missingness must be varied. Remember that missingness is only ignorable under MAR when the parameter of the data is distinct and a-priori independent from the parameter of the missing data process. Under MAR missingness we assume that we may use the observed data to make inferences about the joint (observed and unobserved) data. The dependency of the procedure on the assumption under which we obtain inference is only influenced by the amount of missingness. If there is no missingness–or if there is no data, for that matter–the inference does not depend on the assumption. Alternatively, the validity of assumptions become increasingly important when the missingness increases. Since we control the MAR mechanism, the assumption under which we may solve the missing data problem should hold and it is only fair to assess performance under stringent missingness conditions. We therefore propose, for all mechanisms, to evaluate imputation methodology under 10%, 25% and 50% univariate missingness see Table 2.

Although we limit the focus here to ignorable non-response, the suggested proportions are equally applicable to simulations under non-ignorable non-response.

3.4 Apply missing data method(s) and analysis model(s)

[Use CCA as benchmark method: the imputation method should outperform CCA. Choose imputation model parameters carefully (number of imputations, number of iterations if applicable, predictor matix, regularization if applicable, etc.). Estimate estimand(s) using the analysis model, use correct pooling rules.]

It is wise to evaluate the performance of complete case analysis (aka list-wise deletion) in all simulated conditions. We know the theoretical properties of complete case analysis, which makes the technique useful as a point of departure when evaluating multiple imputation performance.

3.5 Evaluate imputations

Convergence of the algorithm: Most contemporary imputation techniques rely on iterative algorithms, such as the Gibbs sampler, where some algorithms are critically considered to be possibly incompatible Gibbs samplers (PIGS, Li et al., 2012). The convergence of all iterative algorithms should always be considered and if non-convergence is suspected, the inference resulting from the

Table 3: Suggested univariate missingness proportions.

Prop.	Reasoning for suggested missingness proportion
10%	Depending on the size of the data, this percentage can be considered as a lower bound of realistic evaluation. Anything less than 10% may be of little influence on the true data inference. Performance of a missing data method should at least be acceptable for most missing data problems.
25%	This is a fair amount of missingness and will, depending on the observed data information, have a noticeable influence on the completed data inference. When compared to the condition with 10% missingness, the inference obtained under 25% missingness should be less certain (i.e. confidence/credibility interval width should increase), but estimates should still be properly covered and the statistical properties of the missing data method should be sound. In practice, at least to our at least to my experience in social sciences and official statistics, 25% univariate missingness can easily be considered as a realistic missingness percentage.
50%	Performance under 50% percent simulated missingness will most likely be impacted severely. Depending on how the missingness mechanism interacts with the simulated data, some imputation techniques may yield estimates that are under-covered such that the completed data inference should not be deemed valid anymore. If a method yields acceptable inference under 50% MAR missingness, we can determine that the statistical properties of the imputation methodology are sound.

imputations should not be considered. Distributional characteristics: In practice, the distribution of the incomplete data may differ greatly from the observed data. Under anything but the MCAR assumption, this can be expected. When evaluating imputations, the distributional shapes should be checked and diagnostic evaluations should be performed (see Abayomi et al., 2008, for an detailed overview of diagnostic evaluation for multivariate imputations). When anomalies are found, and if the imputation method is valid, there should be an explanation, especially in the controlled environment of a properly executed simulation study.

Fit of the imputation model: [PPC, ref Mingyang, (Nguyen et al., 2017; Zhao, 2022).]

Plausibility of the imputed values: Plausible imputations–imputations that could be real values if they had been observed–are not a necessary condition for obtaining valid inference. However, in practice, especially when the imputer and the analyst are different persons, plausibility of imputations may be a desired property. One would prefer an imputation technique to yield both valid inference and plausible imputations. It should be studied if an imputation method is prone to deliver such impractical results, and if so, under what conditions. When evaluating imputation routines, the evaluator should mention whether the routine is prone to deliver implausible values.

3.6 Evaluate performance

The evaluation criteria may depend on the estimand. When descriptive statistics are the goal and when statistical inference would not be of interest, bias of the estimates would still apply, but standard errors are generally ignored. Chances are that for those who focus on descriptive statistical applications, multiple imputation would not be the mode of choice.

In general, we would say that each multiple imputation routine should be evaluated on at least the following points:

Bias: Results should preferable be unbiased. However, the way bias is considered can greatly influence the interpretation of the results. For example, a negligible absolute bias for a parameter for which the true value is zero, would yield infinite bias when relative bias is considered. The way bias is computed should therefore be carefully chosen and described.

Coverage: Coverage of a 95% interval should in theory be ≥ 95, where a coverage of 95% would be most efficient. Under-coverage (when estimation is too liberal) may be an indication of invalid inference, while over-coverage (when estimation is too conservative) tells us that efficiency could still be gained.

Interval width: The way the confidence interval is calculated should be described. Wider intervals are associated with more uncertainty and the more narrow interval that is still properly covered indicates a sharper inference. However, inference from a wider interval that is properly covered is to be considered more valid when compared to a more narrow interval that is not properly covered anymore.

RMSE: [RMSE at cell level and at prediction level (if applicable).]

Last, when the evaluator is on the verge of drawing conclusions about the performance of the imputation routine, the performance should be carefully qualified. Comparing the performance of an imputation routine given a population (or true) parameter allows for quantitative evaluation. However, in order to pose qualitative statements about the performance on simulated conditions, comparative methodology is required. For example, when claiming that imputation performance is unacceptable when deviations from normality become rather stringent, such performance is highly dependent on the simulation conditions that are used. For a well-balanced judgement about the severity of the performance drop, comparative simulations with e.g. nonparametric models should be executed. A method may perform badly, but if it still outperforms every other approach, it may yet be of great practical relevance.

3.7 Report

Be explicit in choices, better even to share the script directly.

Describe the simulation textually, use pseudo-code, a flowchart, visualize missingness and simulation results.

The used missingness mechanism should be detailed, either graphically or written as a function of the data.

The problem of spurious MAR is generally overlooked in simulation studies and the procedure to generate missing data is often insufficiently described in resulting publications.

Fill in the following checklist.

1. set-up
- aim
- number of simulations
2. data generation
- model vs design based
- data types
- sample vs population (vs bootstrapping??)
- number of observations
- number of variables
- coherence between variables
3. missingness generation
- monotone/univariate vs multivariate
- missingness mechanism
- type of M(N)AR
- missingness proportion

4. missing data methods
- imputation method
- imputation model
- number of imputations
- number of iterations
5. performance evaluation
- analysis model
- estimand
- diagnostic
- ppc

4 Discussion

This document is aimed at establishing a common ground for the evaluation of imputation routines. Such a common ground would be the basis of a standardized evaluation. This allows for fair and efficient comparisons between imputation techniques. Ultimately, it would be desirable to evaluate every imputation routine against the same standardized set in order to quantify the statistical properties across imputation routines. If properly executed, this would allow for careful matching of imputation methodologies to new missing data problems.

Acknowledgements [Maybe add Gelman here?]

Conflict of Interest
The authors have declared no conflict of interest. (or please state any conflicts of interest)

References

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. Journal of the Royal Statistical Society: Series C (Applied Statistics), 57(3):273–291.

Brand, J. (1999). Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. PhD thesis, Erasmus University Rotterdam.

Buuren, S. V., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76(12):1049–1064.

Li, F., Yu, Y., and Rubin, D. B. (2012). Imputing missing data by fully conditional models: Some cautionary examples and guidelines. Duke University Department of Statistical Science Discussion Paper, 1124.

Little and Rubin (2020). Statistical Analysis with Missing Data, Third Edition | Wiley Series in Probability and Statistics.

Little, R. J. (1988). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics, 6(3):287–296.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14):1355–1360.

Liu, D., Oberman, H. I., Muñoz, J., Hoogland, J., and Debray, T. P. A. (2021). Quality control, data cleaning, imputation. In Clinical Applications of Artificial Intelligence in Real-World Data.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9(4):538–558.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(2):371–388.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). Handbook of Missing Data Methodology. CRC Press.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11):2074–2102.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97(4):558–625.

Nguyen, C. D., Carlin, J. B., and Lee, K. J. (2017). Model checking in multiple imputation: An overview and case study. Emerging Themes in Epidemiology, 14(1):8.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics, 19(1):1.

Rubin, D. B. (1976). Inference and Missing Data. Biometrika, 63(3):581–592.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics, 4(1):87–94.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics. Wiley, New York, NY.

Shara, N., Yassin, S. A., Valaitis, E., Wang, H., Howard, B. V., Wang, W., Lee, E. T., and Umans, J. G. (2015). Randomly and non-randomly missing renal function data in the strong heart study: A comparison of imputation methods. PloS one, 10(9):e0138923.

Van Buuren, S. (2012). Flexible Imputation of Missing Data. CRC press.

Vink, G. and van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population.

Zhao, Y. (2022). Diagnostic checking of multiple imputation models. AStA Advances in Statistical Analysis.

