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

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 sub-optimal use of multiple imputation in practice. Additionally, we suggest a course of action for simulating and evaluating missing data problems.

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 policymaking, 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 and data scientists are embracing imputation techniques. After all, the principle of imputation is very intuitive.

The idea behind imputation is to impute (fill in) missing values, 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, p. 76). 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. 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 complete case analysis.

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

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

or more missing data methods. However, no golden standard has been established to evaluate imputation routines and, as a result, the validity and aim of the simulation set-ups 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 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 regarding 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 into the following three distinct categories: problems with data generation, problems with missingness generation, and problems with performance evaluation. We further detail the impact each of these problems may have on the validity of evaluations.

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 generating 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 registers.

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 would subsequently be induced, the observed parts of the data will 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 generalizability 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 some missing entries, for which the true underlying missing data model is–by definition–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 like 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 missing data model. Only sampling complete cases from the data set may thus result in samples that fail to capture all relevant real-world conditions, which in turn refutes the main reason for using design-based simulation. 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. Just to be clear: leaving the missingness as-is and inducing additional missing values to impute could not work here, because we wouldn't have a real and unbiased comparative truth.

After deciding upon a data generating mechanism, another influential–but often overlooked– decision awaits. Simulation studies on missing data methodology have the unique option to exclude sampling variance from their evaluations, and only use the missing data generation procedure as source of Monte Carlo variation. After all, we are interested in the missingness and are not considering the noise induced by the sampling mechanism for evaluation in such studies. Therefore, it is not necessary to take sampling variation into account 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 (see for a detailed discussion Vink and van Buuren, 2014). The only required change to simulation set-ups is to remove sampling variance from performance evaluation. A lot of published work does not consider this option, while it could have sharpened inconclusive or obfuscated results.

2.2 Missingness generation problems

Evaluating imputation methodology requires a missing data problem to be solved. After establishing a comparative truth from a data generating mechanism, some form of missingness therefore 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 mechanisms and missingness patterns. Missingness mechanisms describe the relationship between missing entries and observed data values and are generally categorized into MCAR, MAR and MNAR mechanisms (Rubin, 1976, see Table 1). [TODO: maybe remove mechanisms here?] Missingness patterns concern the location of missing entries across incomplete data (Little and Rubin, 2020, p. 8). Under this definition, there is a row-wise element to the missingness pattern describing which

variables are jointly observed, and a column-wise element encompassing the amount of missingness in the data.

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. The distinction between the two concepts, however, is not always clear in the literature, and convoluting the terms may lead to incorrect generalizations because they rarely mean the same (e.g., a proportion of 50% incomplete cases in bivariate data could translate to a missingness proportion of 25% in both variables, or one completely observed variable and one with 50% missingness). Moreover, the value of the actual missingness proportion may be diffusing too. Some studies use only 10% missingness where other studies push the limits to additionally investigate the performance under missingness proportions of at least 50%. The inconsistent display of simulation conditions may impact the objectivity of meta-evaluations over imputation methods, as one method's performance may appear to be favorable because of 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.

Another aspect of missingness patterns that may misguide evaluations is the complexity of the patterns across variables. It is hardly reasonable to imagine empirical data with only one incomplete variable, yet some simulation studies rely on univariate missingness patterns anyways. Extrapolating simulation results from a specific missing data pattern to more intricate empirical missingness patterns could lead to sub-optimal advice in practice. For example, if a simulator induces missingness in the outcome variable of their analysis model exclusively, they may conclude that list-wise deletion outperforms their imputation method(s). Such a conclusion would, unfortunately, only translate to data that adhere to the same special case (described in Van Buuren, 2012, §2.7), while biasing inferences in other cases. Another example of a missingness pattern that may inadvertently impose assumptions on generalizations is a monotone pattern (i.e. a pattern with uniformly increasing missingness proportions along variables; Little and Rubin, 2020). If a simulator investigates iterative imputation methods under monotone missingness, they might believe that the evaluated imputation algorithms converge instantly, while these methods actually require iteration to produce valid results in all other situations. A final example is multivariate missingness that is generated using step-wise univariate missingness induction. The resulting incomplete data may then not have the desired statistical properties (e.g. skewness; Schouten et al., 2018). This complicates any definitive conclusions about imputation method performance in relation to the data generating mechanism.

Missingness mechanisms are usually assumed to be random (MAR; see Table 1 for definitions). However, the missingness mechanisms that are induced in simulation studies do not necessarily match any true missingness generating mechanisms in incomplete empirical data, nor the mechanisms that are typically assumed when imputing such data. On the one hand, there is MCAR missingness, which may be unlikely as the true missingness generating mechanism, but would yield insightful evaluations in a simulation setting. On the other hand, there is MNAR, which is arguably a quite reasonable source of many real-world missingness scenarios, but is generally ignored except for those evaluations that are specifically targeted at non-ignorable applications. A disconnect between induced missingness, real missingness, and assumed missingness may result in simulation studies that are not as informative as they could be.

MCAR is a necessary simulation condition for evaluating the performance of imputation procedures, yet often omitted. Under MCAR, the statistical properties of the observed data given the missing data are known. 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

Table 1: Missingness mechanisms. [TODO: add PMAR?]

Mech.	Interpretation and consequences of the mechanism
MCAR	Missing completely at random: 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	Missing at random: 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	Missing not at random: The probability to be missing depends 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).

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 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.

Although MAR missingness is often considered as simulation condition, the problem of spurious MAR is generally overlooked. With MAR missingness mechanisms, observed relations in the data are used to induce missingness during simulation (e.g. weight is made incomplete based on observed gender to emulate a situation wherein one gender is less likely to disclose their weight). These 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. The most extreme example would be when 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 pointless.

Inducing a MAR mechanism presents the simulator with another choice, namely the type or functional form of the missingness model. Given the simulated data distributions, one random missingness model may be far more disastrous to the observed information than another model (Schouten and Vink, 2021). 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.

Finally, 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. As stated, this mechanism is not usually included in evaluations.

2.3   Evaluation problems

The evaluation criteria used to assess imputation performance vary from one simulator to another. This is not surprising as people from different fields could have a different focus on the problem at hand. There are, however, some over-arching issues with assessing imputation method performance. In the first place, evaluations of the imputation-generating process (e.g., an iterative imputation algorithm) are generally left out of simulation study results. In the second place, there are performance measures pitfalls to consider.

Many contemporary imputation techniques rely on iterative algorithms, such as the Gibbs sampler, to generate imputations. As with any iterative algorithm–but especially with imputation algorithms that are critically considered to be possibly incompatible Gibbs samplers (PIGS, Li et al., 2012)–algorithmic convergence should be carefully evaluated. Unfortunately, there is no universal quantitative method to diagnose non-convergence in iterative imputation algorithms (Zhu and Raghunathan, 2015; Oberman et al., 2021) and the alternative (visual inspection of the imputation algorithm; Van Buuren, 2012) is neither efficient nor failproof. As a result, imputation algorithms may be terminated before reaching a stable state, which could yield sub-optimal imputations and under-estimated performance of the method. Moreover, failing to produce stable imputations is not an exclusive quality iterative aimputation algorithms. For example if a method does not yield any imputations at all in a certain simulation repetition (e.g. due to over-parameterization), this can be seen as failure in the imputation-generating process too.

The evaluation of simulated results relies on the relation between missing data models, imputation models, and analysis models (i.e. the complete data model used to estimate a target or estimand). When an imputation model is able to capture the essence of the true non-response mechanism relative to the analysis model, the models are said to be congenial (Meng, 1994). Congeniality is difficult to diagnose, but ignoring it may yield sub-optimal imputations. Methods to asses the suitability of imputation models and diagnose misfit often rely on visual inspection of the imputations (see e.g. Abayomi et al., 2008; Bondarenko and Raghunathan, 2016), which is an infeasible endeavor in many simulation studies. Therefore, the performance of imputation procedures on distributional properties is often ignored in simulation studies. 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.

If the goal is inference–not prediction–then the uncertainty about estimates needs to be quantified. A single imputation would then not suffice. The aim of multiple imputation is not to reproduce the data, but to allow for obtaining valid inference given that the data are incomplete. 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). Merely focusing on distance measures would open the door for invalid inferences.

Some simulators may focus on retrieving the unobserved values, while others study the performance in the context of obtaining efficient predictions or classification. For many outcome-driven imputation task a single valid imputation may suffice. Measures of predictive accuracy or the RMSE (root mean squared error) of predicted values could then serve as sensible evaluation criteria.

Table 2: Steps to consider in imputation simulation studies.

	Step	Simulation aspects under consideration
1	Set scope	Aim(s), missing data method(s) to evaluate, simulation design
2	Obtain truth	Data generating mechanism(s), estimand(s), sampling variance
3	Induce missingness	Missingness mechanism(s), missing data pattern(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, qualitative judgement
7	Report	Text, visualization, checklist, simulation script

3 Suggested course of action

Standardizing the evaluation of imputation methodology requires simulators to consider a couple of key aspects in their simulation workflows. Table 2 outlines some suggested steps to adopt. Please note that there is–of course–no 'silver bullet' and that we do not claim to present a universally applicable approach. Our suggestions are the accumulated result of scientific literature, research experience, and discussion. For an excellent overview of general best practices for method evaluation by means of simulation, see Morris et al. (2019). We recommend adhering to their proposed ADEMP structure (aims, data-generating mechanisms, estimands, methods, performance measures) for planning and reporting simulation studies, but highlight and add some specific recommendations for the evaluation of imputation methodology.

3.1 Set scope

Before setting up a simulation study, the simulator should clearly define the scope of their evaluations. A simulation study aimed at imputation method comparison will typically have a different design than one aimed at establishing the inferential validity of a single (novel) method. The choice of imputation method(s) under evaluation is naturally intertwined with the simulation aims and set-up.

One specific simulation study aim (that we have thus far ignored) is to evaluate the predictive performance of pairs of imputation methods and estimation methods in empirical, incomplete data. These kinds of simulation studies do not start out from a complete data set in which missingness is induced by the simulator. Rather, the evaluations use the observed values of an outcome variable (or target) as the estimand, which is estimated from the incomplete predictor (or feature) space by first imputing the missingness and then applying a prediction method. Pairs of imputation and prediction methods with high predictive accuracy in one or more benchmark data sets are deemed as 'good' (Liu et al., 2021). Note that such a simulation design does not have a 'ground truth', so only the comparative performance of methods may be established. We do not recommend this approach if the inferential validity of the imputations is of interest.

Some general choices in the simulation set-up to consider at this stage are the simulation design and the number of simulation repetitions. The simulation design should ideally be fully factorial, i.e. varying each simulation condition against all other conditions. There are, for example, known interaction effects between missingness mechanisms and missingness patterns (the validity of the assumed missingness mechanism becomes increasingly important with higher missingness proportions). If there is more than one imputation method under evaluation, the simulation design should apply each method to every incomplete data set. Applying all methods to the same incomplete data set is computationally convenient, and minimizes unnecessary variation,

Table 3: Source of the comparative truth in simulation studies.

	With sampling variance (multiple samples drawn)	Without sampling variance (one sample drawn)
Model-based simulation	data-generating model	single sample
Design-based simulation	sufficiently large data set	single sample

which makes for fairer comparisons. The number of simulation repetitions may be informed by the required level of precision in the simulation study (e.g. as determined from a maximum tolerable level of uncertainty in terms of a performance measure's Monte Carlo error Morris et al., 2019).

3.2 Obtain truth

The simulator should choose their data generating mechanism(s) in line with the study scope. Model-based data-generating mechanisms have advantages in flexibility and precision, since data are generated from a known statistical model and the true theoretical parameters can be derived. A design-based approach is often used in situations where a probability distribution is not available, or where real-life data structures are of interest.

Estimand(s) or other simulation targets should be defined in the context of the study aim and data-generating mechanism. As a general rule, it would be wise to include one or more multivariate estimands. It is a realistic expectation for imputation methods to preserve both marginal and conditional distributions with respect to the comparative truth. Imputation methods that fail to do so, should not be considered general purpose methods (e.g., mean imputation).

The comparative truth of the estimand(s) is determined by the choice of data generating mechanism and whether sampling variance is included in the simulation set-up, see Table 3. If sampling variance cannot be omitted from the simulation scheme, multiple samples should be drawn from the data generating mechanism (i.e., one sample per simulation repetition, a standard simulation set-up). If sampling variance is not of interest, a single complete data set can be obtained from the data generating mechanism (i.e., one sample for all simulation repetitions). This process is computationally convenient because only a single complete data set has to be considered during all of the simulations. We highly value the additional flexibility that such a simulation set-up offers. Especially in the case of data transformations it can be challenging to derive the true parametric references to evaluate against. Simply using a single generated complete set in which missingness is induced, and evaluating against that generated data set avoids a plethora of procedural problems and computational challenges. When deviating from the standard simulation workflow, conventional pooling rules for multiple imputation (cf. Rubin, 1987, p. 76-77) do not apply. Instead, alternative pooling rules need to be used (Raghunathan et al., 2003; Vink and van Buuren, 2014).

3.3 Induce missingness

Missingness should be induced according to several sets of missing data conditions. We encourage simulators to consider different missingness mechanisms (including missingness types), and different missingness patterns (including missingness proportions). Although not every missingness mechanism is realistically assumed in practice, they can all offer valuable insights as simulation condition.

First, one should always consider MCAR missingness (see Table 1 for definitions). This mechanism should be used as a reference condition. Every imputation method should yield valid imputations under MCAR. We also know that, under MCAR, the distribution of imputed values

should be equivalent to the distribution of observed values for a given variable, which may offer an indication of imputation model (mis)fit. [TODO: leg uit!]

Next, missing data should be induced conform a model that is dependent on the observed data (i.e. a MAR mechanism). Arguably, MAR is the most often-assumed mechanism in practice, and should thus be evaluated carefully. A straightforward technique for inducing univariate MAR missingness is described in Van Buuren (2012, p. 63), generalizations to multivariate MAR missingness can be found in Schouten et al. (2018). If the missingness is to be induced in longitudinal data autoregressive MAR models can be useful (see e.g. Shara et al., 2015, model 2 and model 3) can be useful. 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. The effects of different types of MAR mechanisms are described in Schouten and Vink (2021).

Third, a non-ignorable or MNAR mechanism might fall outside many studies' scope, yet could yield informative insights for daily practice. Even if an imputation method is specifically developed with ignorable missingness in mind, chances are that the method will be applied to non-ignorable empirical missingness at some point in time. It would, therefore, be wise to include MNAR missingness in the simulation study just in case. Alternatively, sensitivity analyses may provide an indication of the validity of the obtained inference, given that the assumed missingness mechanism is suspected to be invalid (see e.g. Molenberghs et al., 2014, part 5).

Apart from the missingness mechanisms, the missingness pattern must be varied. [TODO: beter bruggetje?] 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 to evaluate imputation methodology under 10%, 25% and 50% univariate missingness (see Table 4).

3.4 Apply methods

The imputation methods under evaluation should be applied to the generated incomplete data, after which an analysis model may be fitted to estimate the estimand(s). In addition, the simulator should always perform complete case analysis. We know the theoretical properties of complete case analysis, which makes the technique useful as a point of departure when evaluating imputation performance. Complete case analysis may therefore serve as a benchmark method against which performance should be evaluated. Moreover, pairing complete case analysis with MCAR and MAR mechanisms allows the simulator to evaluate the validity of the missing data generation process under the chosen analysis model.

3.5 Evaluate imputations

[Some suggested minimal requirements for imputation methods: the absence of algorithmic non-convergence, the imputation model fits the observed parts of the data (PPC), under a MCAR mechanism the imputed values are distributed equivalent to the observed values.]

The absence of non-convergence in the imputation models is a minimum requirement for imputation methods. Most contemporary imputation techniques rely on iterative algorithms, such as the Gibbs sampler, where some algorithms are critically considered to be possibly incompatible

Table 4: 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.

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 imputations might be invalid.

[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 a 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.

The fit of the imputation model may be verified with the help of a posterior predictive check (PPC Nguyen et al., 2017; Zhao, 2022). A PPC in the context of imputation methodology could be the multiple over-imputation of observed data values. The reasoning is that if the over-imputed values have equivalent statistical properties to the true data values, the imputation model fits the observed part of the incomplete data well. We might then assume that the imputation model would be able to produce good imputations for missing values too.

[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, plausible 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

[Some suggested minimum requirements for imputation methods: the method yields valid inference under MCAR, the method performs equally well or better than complete case analysis.]

[The evaluation criteria may depend on the estimand. When descriptive statistics are the goal and when statistical inference would not be of interest, the 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: (i) 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. (ii) Coverage: Coverage of a 95% interval should in theory be ≥ 95, where a coverage rate 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. (iii) 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.

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 judgment 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.

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. [TODO: check waar dit al eerder staat!]

3.7 Report

[Be explicit in choices. Describe the simulation textually, use pseudo-code, a flowchart, visualize missingness and simulation results. For example, the used missingness mechanism should be detailed, either graphically or written as a function of the data. And the procedure to generate missing data is often insufficiently described in resulting publications. Fill in the following checklist (in appendix).]

4 Discussion

This manuscript aims to establish a common ground for the evaluation of imputation routines. Such a common ground would be the basis of a standardized evaluation. This would allow for fairer and more 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.

Conflict of Interest
The authors have declared no conflict of interest.

Appendix

A.1. Checklist for the evaluation of imputation methodology

1. Set-up
   - aim
   - target/estimand + analysis model
   - precision level + number of simulations
   2. Data generation
   - model vs design based
   - sampling variance
   - number of observations
   - number of variables
   - data types + structures (e.g. clustering/time-series)
   - coherence between variables
   3. Missingness generation
   - missingness mechanism + type (if applicable)
   - missingness patterns (rows = univariate/multivariate; columns = proportion)
   4. Missing data methods
   - imputation methods + list-wise deletion
   - imputation method parameters (e.g. number of imputations)
   - imputation model parameters (e.g. predictor matrix)
   5. Imputation evaluation
   - imputation process (e.g. non-convergence)
   - imputation model fit (PPC)
   - distribution of imputations (e.g. plausibility)
   - imputation accuracy (e.g. RMSE at cell level)
   6. Performance evaluation
   - statistical properties (e.g. confidence validity)
   - comparative performance (e.g. RMSE of predictions)
   7. Report

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.

Bondarenko, I. and Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. Statistics in Medicine, 35(17):3007–3020.

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.

Oberman, H. I., van Buuren, S., and Vink, G. (2021). Missing the Point: Non-Convergence in Iterative Imputation Algorithms.

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.

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

Schouten, R. M. and Vink, G. (2021). The Dance of the Mechanisms: How Observed Information Influences the Validity of Missingness Assumptions. Sociological Methods & Research, 50(3):1243–1258.

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.

Zhu, J. and Raghunathan, T. E. (2015). Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. Journal of the American Statistical Association, 110(511):1112–1124.