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 methodology. 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 imputation in practice. Additionally, we suggest a course of action for simulating and evaluating missing data problems. [TODO: explain how the standardisation leads to more neutrality, but that this is just a set of elements to consider/report, not a complete 'cookbook'.]

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

Supporting Information for this article is available from www.gerkovink.com/evaluation.

1 Introduction

[TODO: Narrowing the scope of the manuscript to the evaluation of methodology for statistical inference, not prediction or causal inference (although inferential validity encompasses predictive validity and may thus be generalized?). Emphasizing that the ideal evaluation of imputation methodology depends on the aim of the specific study (i.e. how the evaluated methods are used in practice).]

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 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.

[TODO: add that if the goal is statistical inference, the uncertainty due to missingness should be quantified/captured by the missing data method. Refer to Sperrin et al. (2020) for difference in context of prediction. And paraphrase "the quality of a technique can only be evaluated with

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

respect to the aims of the problem/s it's intended to solve in practice". we focus on comparative sim studies]

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 or more missing data methods. However, no gold standard has been established to evaluate imputation routines and, as a result, the validity and aim of the simulation setups 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 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 using simulation studies to evaluate imputation techniques for 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. [TODO: explain how the standardisation leads to more neutrality, but that this is just a set of elements to consider/report, not a complete 'cookbook'.]

Standardizing the evaluation of imputation methodology requires simulators to consider a couple of key aspects in their simulation workflows. Table 3 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.

[TODO: (?) pawel pre-reg simulation protocols? per-protocol results]

2 Why some evaluations should not be trusted

[TODO: Split up the problems with the simulation setup/evaluation and the problems with the interpretation/extrapolation from these simulation studies. See e.g. Greenland (2017) on cognitive problem of generelization versus methodological problems.]

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 Study scope

[Problem: parameters unclear, extrapolation beyond scope of simulations, misinterpretation (see "investigator bias" Greenland, 2017).]

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 setup.

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 setup 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, 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).

2.2 Data generation

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 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 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 is no option here, because we would not 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 the 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, taking sampling variation into account is neither necessary for obtaining information about a method's ability to handle the missing data problem, nor for objectively comparing methods on their ability to correct for missingness (see for a detailed discussion Vink and van Buuren, 2014). The only required change to simulation setups is to remove sampling variance from performance evaluation. Still, a lot of published work does not consider this option, while it could have sharpened inconclusive or obfuscated results. [TODO: add nuance because this doesn't always hold, see Tims review!]

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. [TODO: make this part about the study aim and refer to Petersen and van der Laan (2014) about clear and unambiguous estimand, and to focus on this estimand.] It is a realistic

Table 1: 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

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 setup, see Table 1. 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 setup). 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 setup 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).

2.3 Missingness generation

[TODO: change this to 'be clear which mechanisms are of interest, how they are generated, if they are suitably generated']

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 2). 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

Table 2: Missingness mechanisms. [TODO: add refs Seaman et al. (2013); Mealli and Rubin (2015); Doretti et al. (2018) for clarification and alternatives outlined in Mohan and Pearl (2021); ?); Moreno-Betancur et al. (2018)]

	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.	
v-MAR	Conditional on the fully observed variables missingness occurs at random	(Mohan and Pearl,
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).	

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 whereas 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 while biasing inferences in other cases (Van Buuren, 2018, §2.7). 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 which 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 2 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 is unlikely to be 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 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 a 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 imputation (see e.g. Van Buuren, 2018, §2.7): 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.

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 2 for definitions). This mechanism should be used as a reference condition. Every imputation method must be able to yield valid imputations under MCAR, both in terms of distributional characteristics as well as statistical inference.

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 (2018, §3.2.4), 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). 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, a sensitivity analysis 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).

In addition to varying the missingness mechanisms, each simulated mechanism should be combined with different missingness patterns. 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 becomes 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 several missingness proportions to amulate a realistic range in severity of the missingness problem. Depending on how the missingness mechanism interacts with the simulated data, higher missingness proportions may yield biased or invalid completed data inferences. The missingness proportions should be considered carefully.

2.4 Imputation 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 imputation performance should be evaluated. Moreover, pairing complete case analysis with MCAR and MAR mechanisms allow the simulator to evaluate the validity of the missing data generation process under the chosen analysis model.

2.5 Evaluation

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 overarching issues with assessing imputation method performance. In the first place, diagnostic evaluation of the generated imputations is often left out of simulation study results. Problems with the imputation-generating process (e.g., an iterative imputation algorithm) may then be overlooked. In the second place, there are pitfalls in the evaluation of the estimates that are obtained by fitting the analysis model (i.e. the complete data model used to estimate the comparative truth). The choice of performance measures might inadvertently influence simulation results.

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, 2018) 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. Problems with producing stable imputations is not an exclusive quality iterative imputation algorithms. Any method may run into failures of the imputation-generating process and subsequently lack results (e.g. due to over-parameterization errors).

Omitting further evaluation of the imputation-generating process may yield sub-optimal imputations. The evaluation of simulated results relies on the relation between missing data models, imputation models, and analysis models. 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). [TODO: change this to being able to embed the analysis procedure and the imputation model within a Bayesian model; the nonresponse mdel may be misspecified while the imputation procedure is congenial.]The congeniality of all imputation models should be assessed, but methods to assess 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). This makes its assessment an infeasible endeavor in many simulation studies. Moreover, the performance of imputation procedures on distributional properties is often ignored too. 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.

The second source of problems with evaluating simulation outcomes is the choice of performance measures. Performance measures should express how well the obtained estimates approximate the comparative truth, but not every metric is suited for the evaluation of imputation methodology. For example, using measures of predictive accuracy or the RMSE (root mean squared error) as the evaluation criterion may increase the rate of false positives (Van Buuren, 2018, §2.6). And, under certain simulation conditions, an otherwise useful metric such as relative bias can distort the interpretation of results (i.e., a negligible absolute bias may yield an infinite relative bias if the true value of the estimand is zero). The usefulness of evaluation criteria may thus depend on the simulation goal and estimand.

If the goal is inference–not prediction–then the uncertainty about estimates needs to be properly quantified. To capture this uncertainty, the standard errors of the estimates should be correctly

calculated, which requires multiple imputation. 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). [TODO: add nuance about coverage rate $> 95\%$ irt CIW and efficiency.]

The choice of performance measures may inadvertently distort statistical properties of the imputed data. Developers often only inquire about the 'accuracy' (i.e. how well can the method reproduce the original data). Merely focusing on the discrepancy between the true and imputed data would open the door for invalid inferences.

Every imputation workflow should contain an evaluation of the obtained imputations. Even though inspecting each imputation may be labor-intensive due to the number of imputations generated in a simulation study, we highly recommend to consider the following aspects:

(i) The absence of non-convergence in the imputation-generating process is a minimum requirement for any imputation method. If non-convergence is suspected, the inference resulting from the imputations might be invalid. However, preliminary work suggests that iterative imputation algorithms could achieve inferential validity before reaching a stable state (Oberman et al., 2021).

(ii) The fit of the imputation model may be verified with the help of a posterior predictive check (Nguyen et al., 2017; Zhao, 2022). A straightforward posterior predictive check for imputation methodology is the multiple over-imputation of observed data values. If the statistical properties of the over-imputed values are equivalent to those of the observed data values, one could infer that the imputation model fits the observed part of the incomplete data reasonably well. Then, by extension, one could assume that the imputation model might be able to produce good imputations for the missing part of the data too. [TODO: add that this is implemented in amelia.]

(iii) The distributional characteristics of the imputations should be inspected for anomalies. 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.

(iv) Finally, the plausibility of the imputed values may be evaluated. 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. [TODO: add nuance here because valid inference, not plausible imputed values should be the aim.]

Imputation method performance should be quantified using appropriate measures. It depends on the specifics of each simulation study which performance measures would be most suitable. In general, we recommend to evaluate at least the following points:

(i) The estimates should preferably be unbiased. Note that the way bias is calculated should be carefully chosen and described, since this can greatly influence the interpretation of the results. Unbiased estimates under MCAR are a minimum requirement for any imputation method.

(ii) The estimates should have a proper coverage rate. 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 and intervals are too narrow) indicates that the procedure is not confidence

Table 3: Steps to consider in imputation simulation studies.

	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 method(s), analytic method(s)
5. Evaluate imputations	Imputation-generating process, imputation model fit
6. Evaluate performance	Performance measures, qualitative judgment
7. Report	Text, visualization, checklist, simulation script

valid and may lead to invalid inference. Over-coverage (when estimation is too conservative and intervals are too wide) tells us that efficiency could still be gained.

(iii) The width of the confidence or credible interval may convey statistical efficiency, which should be considered to compare imputation methods. Wider intervals are associated with more uncertainty whereas a 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 than a more narrow interval that is not properly covered anymore.

We do not generally recommend the use of the RMSE as evaluation criterion, because this metric does not account for the inherent uncertainty of the missing values (Van Buuren, 2018, §2.5). However, if a study is aimed at obtaining predictions or classification and inferential validity is not of interest, the RMSE of the predictions may yield valuable information about the methods' efficiency in terms of both accuracy and precision. [TODO: make this an optional bullet point for prediction only: out of sample mean squared prediction error can be compared between methods.]

Last, when the simulator is on the verge of drawing conclusions about the performance of the imputation methodology, the performance should be carefully qualified. Comparing the performance of an imputation routine given a population (or true) parameter allows for quantitative evaluation. Yet, 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.

2.6 Reporting

After the careful consideration and execution of all simulation aspects, the evaluations should be properly reported. We encourage simulators to document their choices and to be explicit in their descriptions. For example, the simulation design may be presented textually, in a flow chart or as a block of pseudo-code, whereas missingness mechanisms may be written as a function of the data or displayed graphically. Ideally, this should be supplemented by an online repository with all of the data and code required to reproduce the simulation results. To aid simulators in reporting and move towards standardization in evaluation, we provide a draft version for reporting guidelines in Appendix A.1 (also available from www.gerkovink.com/evaluation). We invite the readers of this paper to contribute to its development.

[TODO: split up into design, execution and reporting? (?)]

3 Discussion

This manuscript aims to elicit critical thinking about incomplete data simulation and 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.

Acknowledgements We thank the Amices team for the fruitful discussions.

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

Appendix

A.1. Reporting guidelines

Table 4 provides a checklist for reporting on imputation methodology evaluations.

Table 4: Checklist for reporting on imputation methodology evaluations.

1	Simulation scope
	Aim
	Design (incl. pseudo-code or flow diagram)
	Number of simulation repetitions
2	Comparative truth
	Data-generating mechanism (model-based or design-based)
	Sampling variance
	Data characteristics (incl. multivariate relations and structures e.g. clustering)
	Estimand
3	Induced missingness
	Missingness mechanism (incl. type or functional form of the missing data model)
	Missingness pattern (incl. missingness proportion)
4	Applied methods
	Imputation methods (incl. parameters e.g. the number of imputations)
	Analytic methods (incl. calculation of standard errors e.g. pooling rules)
	Reference method (e.g. complete case analysis)
5	Imputation evaluation
	Imputation-generating process (e.g. algorithmic non-convergence)
	Imputation model fit (e.g. posterior predictive checks)
	Distributional characteristics (e.g. plausibility of imputed values)
6	Performance evaluation
	Statistical properties (e.g. confidence validity)
	Comparative performance (e.g. predictive accuracy)

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.

Doretti, M., Geneletti, S., and Stanghellini, E. (2018). Missing Data: A Unified Taxonomy Guided by Conditional Independence. International Statistical Review, 86(2):189–204.

Greenland, S. (2017). Invited Commentary: The Need for Cognitive Science in Methodology. American Journal of Epidemiology, 186(6):639–645.

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.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika, 102(4):995–1000.

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

Mohan, K. and Pearl, J. (2021). Graphical Models for Processing Missing Data. Journal of the American Statistical Association, 116(534):1023–1037.

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.

Moreno-Betancur, M., Lee, K. J., Leacy, F. P., White, I. R., Simpson, J. A., and Carlin, J. B. (2018). Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. American Journal of Epidemiology, 187(12):2705–2715.

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.

Petersen, M. L. and van der Laan, M. J. (2014). Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. Epidemiology, 25(3):418–426.

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.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What Is Meant by "Missing at Random"? Statistical Science, 28(2):257–268.

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.

Sperrin, M., Martin, G. P., Sisk, R., and Peek, N. (2020). Missing data should be handled differently for prediction than for description or causal explanation. Journal of Clinical Epidemiology, 125:183–187.

Van Buuren, S. (2018). Flexible Imputation of Missing Data. Chapman and Hall/CRC.

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.

