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 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: 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). 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. Explain how the standardisation leads to more neutrality, but that this is just a set of elements to consider/report, not a complete 'cookbook'. Add that we focus on simulation studies evaluating imputation methodology for statistical inference.]

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

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 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 and developing imputation methodology. 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. The completed data that are thus obtained can be analyzed by standard techniques, which conveniently separates the missing data problem from the analysis.

For some data analyses a single imputation may suffice, while for other data analyses multiple imputations are needed. For example, in the case of inferential analyses, such multiple imputations (Rubin, 1976) have proven to be a valuable technique for obtaining valid inferences on incomplete datasets. The resulting multiple inferences on multiple completed data sets can be combined into a single inference using Rubin's rules (Rubin, 1987, p. 76). When generating predicted values is the goal of the data analysis, a single imputation may already generate sufficient predicted values (Nijman et al., 2020).

---

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

In all scenarios it can be said that the quality of the solution obtained by imputation depends on the statistical properties of the incomplete data and the degree to which the 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. Such studies are typically composed of an analysis pipeline in which data are subsequently generated, made incomplete, and imputed repeatedly under different simulation conditions. A set of evaluation criteria is then postulated to evaluate the performance of one or more missing data methods. Recent attention on best practices for method evaluation by means of simulation (Morris et al., 2019). and **REF MISSING** Pawel et al. (2022)'s advice to pre-register simulation protocols will guide methodologist in conceptualizing their simulation workflow. However, there is no consensus on how to design, execute and report simulation studies aimed to evaluate imputation methodology. Without a 'gold standard' for evaluation, the validity and comparability of simulation setups may differ tremendously from one developer to another. This brings forth a chain of potential problems in the objective assessment of imputation method performance within and across studies, which could lead to sub-optimal use of imputation in practice.

The purpose of this paper is threefold: First, we raise some concerns with respect to evaluating imputation methodology. These concerns stem from careful consideration with fellow 'imputers' and from encounters as reviewers for statistical journals. Second, we 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 our 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

The ideal evaluation of imputation methodology may differ between studies. A simulation study developed for the comparison of several imputation methods will typically have a different design than one aimed at establishing the inferential validity of a single (novel) method. How the evaluated methods are used in practice thereby governs the simulation procedure. Since the simulation aims and setup are intertwined with the choice of the evaluated imputation methods, one should always asses the suitablity of the chosen evaluation metrics with the imputation methods' purpose in mind.

We limit the scope of this paper to comparative simulation studies in the context of statistical inference. With that, we exclude the evaluation of imputation methodology for prediction or causal inference. The type of simulation study under consideration in this paper requires a form of 'comparative truth' to assess the inferential validity of imputation methods. This is in contrast to simulation studies aimed at comparing the predictive performance of imputation and prediction method pairs. Such a simulation design typically does not start out from a 'ground truth' (e.g., a complete data set in which missingness is subsequently induced by the simulator). Rather, the imputation and prediction method pairs are evaluated on their ability to yield high predictive accuracy in one or more incomplete benchmark data sets (Liu et al., 2021). In these designs, 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. For an overview of missingness in the prediction context see e.g. Wood et al. (2015) and Sperrin et al. (2020). Missing data within a causal inference framework has been described by e.g. Moreno-Betancur et al. (2018) and Mohan and Pearl (2021).

Within the scope of comparative simulation studies aimed at statistical inference, we observe several problems. To demonstrate the broad impact of these problems, we recognize the following four distinct categories: problems with simulation design, 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 Simulation design

Before setting up a simulation study, the simulator should clearly define and describe the scope of their evaluations. If the simulation parameters are unclear, any methodological problems with simulation study design may be obfuscated by the cognitive problem of generalization (Greenland, 2017). For example, misinterpretation and (unintended) questionable research practices can prompt extrapolation beyond the scope of the simulations (Greenland, 2017; Pawel et al., 2022). Imputation methods that seem to work well in simulations may then not be suitable in practice, when applied on incomplete empirical data.

One potential reason for discrepancies between imputation method performance in simulations versus applications is the use of sampling variance in the simulation design. 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 (Vink and van Buuren, 2014). If sampling variance is not of interest, a single complete data set can serve as comparative truth in all simulation repetitions (as opposed to a 'standard' simulation setup, in which data is drawn every repetition). After all, we are interested in solving for the missingness and have no need for the noise induced by the sampling mechanism for evaluation in such studies.

Removing sampling variance from performance evaluation is computationally convenient and may potentially sharpen comparisons between imputation methods, clarifying inconclusive results on their ability to solve for missingness. However, when deviating from the standard simulation workflow, conventional pooling rules for multiple imputation (cf. Rubin, 1987, p. 76-77) do not apply and alternative pooling rules need to be used (Raghunathan et al., 2003; Vink and van Buuren, 2014). This opens the door towards human error in the implementation and interpretation of such a simulation study. On the one hand, we highly value the additional flexibility that this design 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. On the other hand, we want to warn simulators against potential mishaps, e.g. when calculating and comparing standard errors.

A well-designed simulation does not guarantee fair evaluations. Estimands and other simulation targets should be clearly and unambiguously defined in the context of the study aim Petersen and van der Laan (2014). 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). The simulation design should ideally be fully factorial, i.e. varying each simulation condition against all other conditions. There may, for example, be interactions between the source and amount of missingness in the incomplete data and the efficacy of imputation methods (i.e., the validity of the assumed missingness mechanism becomes increasingly important with higher missingness proportions) [TODO: add ref dance of the mechanisms!]. 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.

2.2 Data generation

[TODO; describe problems with data generation and interpretation: If the source of the data is model-based, the correlations in the data may be so high that there is no effective missingness due to the predictive power in the other variables, and the data source may be unfair if the data generating model is equal to one of the imputation models under evaluation. With a design-based simulation, the data source is unlikely to have no missingness and the choice of initial missing data method may propagate into the evaluations.]

To evaluate the ability of an imputation routine to handle missingness, a form of ground 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 from a sufficiently large observed set, such as official registers. The simulator should choose their data-generating mechanism 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.

The challenge 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 known 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 potentially 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 measurement level, 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 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.

2.3   Missingness generation

[TODO: Make clear in this section that missingness mechanisms can be spurious (spurious MAR when correlations are too low, or when missingness proportion is too low). In reporting, the functional form or type of missingness is often omitted. As advice, MCAR can be a very informative mechnism for reference and MNAR may be useful for practice. Missingness proportions can be misinterpreted. Missingness pattern may have unexpected consequences (e.g., univariately generated multivariate missingness). In interpretation, watch out for generalizations from unintended special cases.]

Often reports of simulation studies remain vague about the 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 whereas 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). Simulators should therefore be explicit in their description of the missing data generating model and should make sure that the model fits the aim of their simulations.

Extrapolating simulation results from a specific missing data generating model to more intricate (empirical) missingness 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 inadvertently induce missingness according to the special case descibed by Van Buuren (2018, p. §2.7). Under this specific missingness model, list-wise deletion may outperform any imputation method. Such a conclusion would, unfortunately, only translate to data that adhere to the same special case, while biasing inferences in other cases. Another example of potential side-effects of missingness generation may occur when multivariate missingness is generated using step-wise univariate missingness induction. The resulting incomplete data may then not have the desired

statistical properties (Schouten et al., 2018). This complicates any definitive conclusions about imputation method performance in relation to the data generation parameters.

Missingness should be induced according to several sets of missing data conditions. We encourage simulators to consider different missingness patterns and mechanisms. Missingness mechanisms were first defined in Rubin's seminal work (Rubin, 1976). For more recent considerations see e.g. Seaman et al. (2013); Mealli and Rubin (2015); Doretti et al. (2018); Moreno-Betancur et al. (2018); Schouten and Vink (2021); Little and Rubin (2020); Mohan and Pearl (2021); Schomaker (2021). Although not every missingness mechanism is realistically assumed in practice, they can all offer valuable insights as simulation condition. A disconnect between induced missingness, real missingness, and assumed missingness may result in simulation studies that are not as informative as they could be.

Truly random missingness across all data entries ('missing completely at random' or MCAR under Rubin's definition) may be considered a necessary simulation condition for the evaluation of imputation procedures, since 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 universally sound. Sadly, the straightforward case of MCAR is often neglected from simulation studies, while it offers an informative reference condition in many simulation setups.

Besides MCAR missingness, data dependent missingness mechanisms, such as missing at random (MAR) missingness are paramount in statistical inference investigations. 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, generating the missingness through 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).

Inducing a MAR mechanism presents the simulator with a 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 functional form of the missingness-generating model.

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 mimic 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 true 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 generated randomly (see Figure 1). 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.
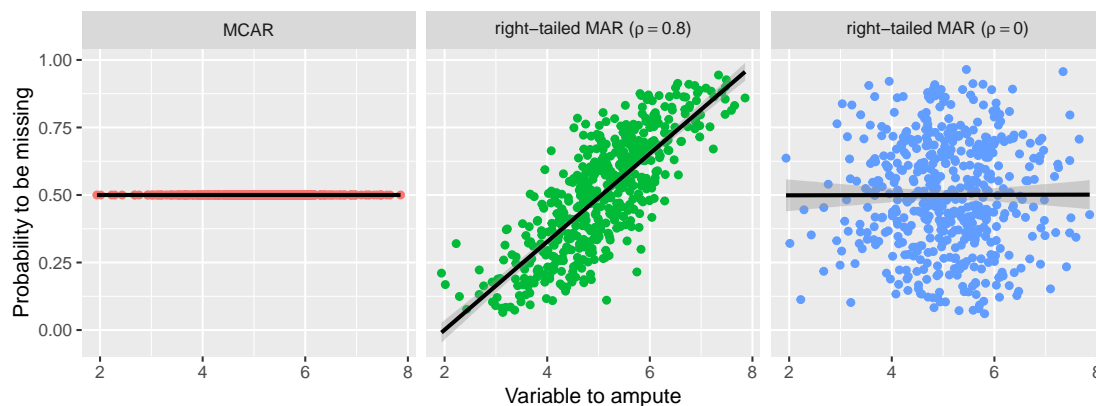
Figure 1: Three different realizations of missingness generation for MCAR and right-tailed MAR missingness mechanisms for different levels of correlation ρ between the variable to ampute and the covariates that guide the probability to be missing. Note that the conditional probabilities of MCAR and right-tailed MAR ($rho = 0$) are identical.

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. After all, for every MNAR mechanism there is a MAR mechanism with equal fit and one cannot definitively verify that empirical missingness is random (Molenberghs et al., 2008). Therefore, it can be argued that MNAR is the more likely mechanism for real-life missingness scenarios. It may 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 (Little and Rubin, 2020, Corollary 6.1A). 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 therefore be considered carefully.

The above emphasizes the need for both thorough evaluation and thorough documentation of the intended and used missingness mechanisms. Simulators must be make clear which mechanisms are of interest. To be able to verify these mechanisms, simulators myst also be transparant about the process that generated the missingness. For example by sharing the code or detailing the

process that generated the missing values. Alternatively, missingness generation devices, such as the mice::ampute() function can be used to generate valid, pattern-based missingness in any structured data set (Van Buuren and Groothuis-Oudshoorn, 2011; Schouten et al., 2018). Finally, simulators must include simulation conditions and evaluation parameters that make it possible to identify the proper generation of missing values. This would prevent that errors in the missingness generation would go unnoticed.

2.4 Performance evaluation

[TODO: make this section less stringent and focus more on the suitability of the performance evaluation in the context of the study aim. The take-away should be that simulation results cannot be easily compared if the performance measures differ. So we suggest to use CCA as benchmark method, and if possible also complete-data inference, to provide a lower and upper limit for expected imputation method performance. And even if the goal of the simulation effort is imputation method comparison, one should not just evaluate relative performance between imputation methods, but also report statistical properties of estimates such as bias, confidence intervals, and the coverage rate of the confidence intervals.]

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, there are pitfalls in the evaluation of the estimates that are obtained by fitting the analysis model after imputation. For example, focusing on one performance measure over another, may yield different conclusions about imputation efficacy. In the second place, diagnostic evaluation of the generated imputations is often left out of simulation study results. Identifying problems with the imputation-generating process (e.g., an iterative imputation algorithm) may offer explanations for under-performance of imputation methods. Such evaluations, however, are typically omitted and valuable insights into the imputation method(s) may then be overlooked.

At minimum, 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. 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. A negligible absolute bias, for example, may already yield an almost infinite relative bias if the true value of the estimand is zero. In most cases, unbiased estimates may be expected under a MCAR missingness mechanism which can be easily verified by including MCAR missingness to the scope of missingness investigations.

(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 (Neyman, 1934). Under-coverage indicates that the procedure is not confidence valid and may lead to invalid inference. Undercoverage may occur when the estimation procedure is either biased, too liberal leading to intervals that are too narrow, or both. Over-coverage, though technically not confidence invalid, would indicate that efficiency could still be gained. In such cases, a narrower interval exists for which the method would still yield confidence valid result.

(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. That said, with valid nominal coverage, the method that yields the narrowest interval would be most efficient.

(iv) Predictive accuracy may be quantified using the root mean squared error (RMSE). 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 and it may increase the rate of false positives (Chapter 2.6 Van Buuren, 2018). However, if a study is aimed at obtaining predictions, and inferential validity is not of interest, the RMSE of the predicted values can be compared between method, which may yield valuable information about the methods' efficiency in terms of both accuracy and precision.

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 always be studied.

GV CONTINUE HERE

Next to imputation method performance, the imputation-generating process may be evaluated. 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. Under many circumstances, imputation methods may be realistically expected 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).

[TODO: add segway to this paragraph.] 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).

[TODO: integrate this into text above!] 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 (although plausibility is nice).]

[TODO: add segway to reference method paragraph and add that performance on the complete data offers an upper limit for performance.] 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.

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.

3   Suggested course of action

After the careful consideration and execution of all simulation aspects, the evaluations should be properly reported. 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.

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.

We aim 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, such evaluations 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. Reporting guidelines

Table 1 provides a checklist for reporting on imputation methodology evaluations.

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.

Table 1: 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 source (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)

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.

Nijman, S. W. J., Hoogland, J., Groenhof, T. K. J., Brandjes, M., Jacobs, J. J. L., Bots, M. L., Asselbergs, F. W., Moons, K. G. M., and Debray, T. P. A. (2020). Real-time imputation of missing predictor values in clinical practice. European Heart Journal - Digital Health, page ztaa016.

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

Pawel, S., Kook, L., and Reeve, K. (2022). Pitfalls and Potentials in Simulation Studies.

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.

Schomaker, M. (2021). Regression and Causality.

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.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(1):1–67.

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

Wood, A. M., Royston, P., and White, I. R. (2015). The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. Biometrical Journal. Biometrische Zeitschrift, 57(4):614–632.

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.