GENERAL/MAJOR COMMENTS

This is a really interesting paper that deals with long-standing issues in missing data research. There are lots of interesting discussion points and proposals. Overall, the piece is well written and I appreciated the careful use of terminology that is so often lacking, which made it very readable. What I felt was lacking was the rationale for several of the views presented, which felt a bit narrow; more below. While I am positive about what you are trying to create here, there is no argument made for how the standardised approach is 'neutral', and some opinions need to be broadened/tempered and better qualified/justified.

* **Thank you for the feedback. We have aimed to add more depth to the revised manuscript.**

While reading the article I kept in mind the following (non-exhaustive) list of situations where imputation is theoretically attractive and/or used in practice:

1. Incomplete covariate values in observational studies with focus on unbiased, efficient parameter estimation; 2. Incomplete covariate values in RCTs with focus on unbiased, efficient parameter estimation; 3. Incomplete outcome values in RCTs or observational studies with focus on unbiased, efficient parameter estimation; 4. Incomplete predictor values for prognosis/prediction tasks.

The views expressed in the manuscript suggests the authors are likely motivated by (1) because the advice tends to correspond to what is relevant in this context. There are several points that would be different for situations (2)–(4); see specific comments. I think it would be worth laying out in the introduction the type of missing data imputation problems this framework targets. Stating a primary focus on (1) alone would be the easiest thing to do while it would require a major re-think to be applicable for situations (2)–(4).

* **We choose option (1).**

There is a recommendation in section 2.1 to remove sampling variation from simulation studies to focus the problem on the missing data. This is certainly a technique that authors have used but it is not a general solution and, in my opinion, should be used only where bias is the sole performance measure. So for example I teach this as a possibility in simulation studies short-courses but with the strong caution that the approach changes the rules for standard errors, coverage, etc. (more detailed example below). The manuscript does not acknowledge the implications or consider the potential for this creates for confusion.

* **We emphasized that this is not a fix-all and we remind the readership about this at multiple occasions, when we feel that this is relevant.**

Section 2.2 discusses missingness mechanisms but without really justifying the views presented. Some points were controversial but some were incorrect (detailed comments below). It is hard to justify some, e.g. saying that the missingness mechanism may not be realistic. As with other parameters, what is 'realistic' for one dataset is not for others but, unlike with other parameters, with missing data the mechanism is not ever observable. I would suggest changing the advice to say that people need to be clear about which missingness mechanisms are of interest, reassure themselves that a mechanism is of suitable 'strength' (e.g. not claiming MAR when it is very close to MCAR), and making sure that an imputation technique is labelled as suitable – or not – only for scenarios that have actually been explored.

* **We adopted the suggestions.**

ABSTRACT

The abstract is clear but rather terse. The main question that arises from reading it is what sort of standardization you are proposing: is this a set of principles to follow or practical hurdles to clear; if the latter, is this a ‘bare minimum’ or something that aims to encompass everything (much harder)?

* **The revised manuscript has a revised abstract that clearly explains the aims of the work and invites the readership to collaborate.**

INTRODUCTION

Very nicely written. I like the sentence ‘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 modelling missing values’ but it misses something. Before anything listed, I think it’s important to say that the quality of a technique can only be evaluated with respect to the aims of the problem/s it’s intended to solve in practice. For example, mean imputation and missing indicator techniques are terrible ideas in general but both are fine (sometimes optimal) for a partially observed covariate in RCTs (White & Thompson, Stat Med, 2005) or where the task is prediction rather than parameter estimation (Sperrin et al., JCE, 2020).

* **Thanks for the suggestion. We emphasize the match between the missing data problem and the evaluation methodology much more in the revised manuscript.**

There is mention of ‘(semi-)parametric’ but this doesn’t really come up later in the manuscript. Does this refer to hot-deck methods? ‘Semiparametric’ has a fairly specific technical meaning in the sense that e.g. people like Tsiatis would use it, which is why Schenker and Taylor (1996) used the term ‘partially parametric’.

* **We agree. We have changed this to `hot-deck` as it is more fitting to the imputation technique.**

Page 2 says, ‘Second, to provide imputation methodologists with a suggested course of action when simulating missing data problems’. This sounds like it’s simply about how data are generated when the actual aims are broader. Should this say ‘…action when using simulation studies to evaluate imputation techniques for missing data problems’? That is, it’s specifically targeting imputation rather than generally missing data simulation studies.

* **We included this suggestion in the revised manuscript.**

SECTION 2

The beginning of this section reminded me of a recent arxiv pre-print from Samuel Pavel and colleagues, ‘Pitfalls and Potentials in Simulation Studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method’. The section begins by saying there is currently no consensus and part of your aim is to work towards some consensus. I think it’s worth first stating arguments against any consensus or concerns. For example, any consensus necessarily introduces a certain amount of rigidity and it’s hard to be broad enough to account for different imputation tasks given different analysis aims. As I write this paragraph, I have not yet read your proposals, but such concerns will clearly need to be taken into account for them to be accepted and used by imputation researchers.

* **We refer to the suggested reference in the revised manuscript.**

Section 2.1 says ‘the luxury position to establish the truth beforehand by choosing a data-generating mechanism’. This presumably refers to what would be targeted with complete data. It is worth noting that this is frequently but not always the target: an important point in Wood et al. (2015, ‘The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data’, Biometrical Journal). They define ‘pragmatic’ performance as performance for a future setting where predictors are not always observed.

* **Thanks. We agree; but we have decided to narrow the scope of the revised manuscript to inferential evaluations and spend little time on the evaluations of predictions after imputation. We highlight some recent references that focus attention on predictions in this context, but feel that more work is needed in order to give a conscientious recommendation of imputation procedure evaluation in the context of prediction.**

Section 2.1 Talks about the issues with model-based simulation of complete data – a fair critique. However, it is worth mentioning that this is also a problem with model-based simulation of missing data, addressed in section 2.2.

**🡪 We have changed the order of appearance in the revised manuscript.**

Section 2.1 the final paragraph discusses removing the complete-data sampling variation so that simulation studies include only sampling variation due to incomplete data. I’ve had lots of discussions about this, including with authors as a reviewer. The key advantage is that the full-data analysis is exactly correct and has no Monte Carlo error, but there are issues that cause people to trip-up. In particular, the rules for assessing standard errors change as follows:

1. Empirical SE. Consider a superefficient multiple imputation method (to make the point, where superefficient is in the sense of Meng 1994 and Rubin 1996 ‘MI after 18+ years’). The empirical SE of the full-data analysis is 0, since it is fixed across repetitions. The empirical SE of the superefficient method is bounded to be positive, since it will vary due to missing data uncertainty, when it is more efficient than the full-data analysis. This cannot be detected by the simulation study, which may miss reasons to be wary of a method.

2. While the average model SEs can be compared across methods, these cannot be compared with the empirical standard errors, which is an important way to evaluate their bias.

3. Similarly, consider coverage. The coverage of the full-data analysis is by definition 100%, when with data-sampling variation we would hope to achieve 100(1-\alpha\%). I have seen researchers using this simulation technique naively assess coverage and cheer when their imputation methods achieve '95%', but this does not correspond to 95% coverage in the usual sense and so coverage should not be assess.

🡪 **We agree. Perhaps our enthusiasm has made our statements a bit less political in the earlier manuscript.** **We have restructured the revised manuscript greatly and added a lot of nuance that address the above points.**

Section 2.2: I enjoyed this section. The important distinction between ‘incomplete cases’ and ‘missing for a specific variable’ is nice and constructive.

* **Thank you!**

Table 1: There are several versions of ‘MAR’ floating around. The classic Rubin ‘patterns’ definition is that missingness depends only on variables when observed (so can depend on values of an incomplete variable when that variable is observed but not when missing), while the definition (confusingly IMO) used by Pearl and Mohan is that missingness can only depend on fully-observed variables. So some things that are Rubin-MAR are Mohan-MNAR. It might be good to mention this when describing the taxonomy and also not which version you are using.

🡪 **The respective table is not in the revised manuscript anymore. We have added nuance and added the Mohan reference.**

In section 2.2, page 4, the second paragraph has several points that are controversial in the sense that they seem to carry the authors’ particular interest in multivariate imputation. For example:

1. ‘It is hardly reasonable to imagine empirical data with only one incomplete variable, yet some simulation studies rely on univariate missingness patterns anyways’. This depends on context; in a study with missing data in the outcome at the primary time, it is perfectly reasonable not to bother checking how an imputation technique would perfrorm with multivariate missingness.

2. ‘missingness in the outcome variable of their analysis model exclusively’. Again, the context is all-important and, if this is the intended use-case of an imputation method, this is sensible.

3. Monotone patterns: I am slightly more sympathetic to this one but the pattern is not unusual in studies with a repeatedly-measured outcome.

**🡪 Again, the revised manuscript is much more nuanced.**

In section 2.2, the penultimate paragraph talks about missingness mechanisms. I suspect there is something unsaid here, which needs clarification. The authors are either considering that simulators should want to demonstrate a MAR-based imputation method has reasonable properties under MNAR, or that MNAR scenarios and implementations of an imputation method should be considered under MNAR. Neither is convincing. ‘Default’ imputation methods are typically intended to be valid under MAR unless some external information is injected. Researchers should be using imputation methods only when they are willing to assume a certain mechanism. If the researcher running the simulation study clearly labels a method as appropriate under MAR but not otherwise, are they obliged to include MNAR adaptations of their method or MNAR mechanisms in their simulation study?

* **We have reformulated our recommendation to investigate MNAR missingness in the evaluations more explicitly as an advice.**

Section 2.2: ‘MCAR is a necessary simulation condition’. Again, this is controversial. Consider a simulation study that evaluates imputation methods adapted for MNAR, e.g. reference-based multiple imputation. Whether it performs satisfactorily under MCAR is irrelevant to the question of whether it is a useful imputation technique with incomplete data. Given that we rarely if ever believe MCAR, why should we care about its performance there?

* **We have nuanced the recommendation in the revised manuscript.**

Section 2.2: ‘one of the special cases under which complete case analysis would be more efficient than imputation’. It’s incorrect to describe this as a ‘special’ as it’s not a subset of MAR. The 'complete-case assumption' is P(M|Y,X)=P(M|X). To be valid, complete case analysis does not require missingness to be independent of X, even if X is missing, and this is not a special case of MAR.

* **TODO: check if the proposed change resolves this reviewer comment**

Section 2.3 – great point about using diagnostics etc. This always concerns me as well but it’s hard to know what to do or advise others to do. Is it realistic to aim to program simulation studies to behave as an analyst would? To what extent? Do the authors have good examples to reference?

* **This section has been restructured to emphasize 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).**

Section 2.3 I think the comment on RMSE is about RMSE of predictions rather than the simulation RMSE and suggest a clarification. What is meant by ‘false positives’ in the subsequent sentence?

* **This term has been replaced by the chance of type I error.**

Section 2.3: The final two paragraphs are really nice and I especially liked that it started, ‘If the goal is inference…’ I wondered about the advice if the goal is prediction?

* **We have narrowed the scope of the manuscript to the evaluation of methodology for statistical inference, not prediction.**

SUGGESTED COURSE OF ACTION

The manuscript came into its own in section 3 and there were lots of constructive, helpful proposals. I thought you were more open and circumspect about things that had come across less well in section 2.

* **Thank you. We have changed the structure of the revised manuscript to combine sections 2 and 3.**

One thing that surprised me was the description of ‘induced missingness’ as distinct from the data-generating mechanism. I really like that there is focus on it but induced missingness is a part of the data-generating mechanism.

* **We have renamed the subsections to ‘data generation’ and ‘missingness generation’ (which are, indeed, both part of the data generating mechanism).**

In section 3.2, I really liked the start that talks about the scope of the study. However, this changes in paragraph 2, where a single estimand is warned against. I would suggest that this should be in line with the previous paragraph. You argued at the end of section 2 that bias, confidence interval width and coverage proportion should be of interest. If these are satisfactory, why should we care about face validity of imputations? The classic example here is mean imputation of incomplete covariates in randomised trials: trials are designed to estimate a treatment effect, not intercepts or covariate effects, so a single estimand is appropriate.

* **We have added our reasoning for the evaluation of imputed values in the revised manuscript.**

I was surprised to see the suggested rule-of-thumb about the proportion of missingness to evaluate table 4. The proportions investigated surely should depend on the context, no? (I grant that people often just say ‘20–30% missing data should be ok’ because it’s a sweet spot for MI: too much missing data for complete case analysis to be ok but not enough for MI to struggle.) Is the real idea here 'look at >1 missingness proportion and consider carefully'?

* **We have substituted the table with specific suggested missingness proportions for more general advice to evaluate several missingness proportions.**

Under ‘apply methods’ it has some very good advice e.g. always include complete case analysis. I wonder if this requires a rationale, e.g. just as full-data analysis usually provides an upper bound for performance, complete-case analysis should be a lower bound for any imputation method to beat.

* **We have adopted the suggestion.**

Section 3.5 points iii–iv talks about evaluating imputations rather than valid inference and I have the same comment as above. Isn’t this similar to comparing imputed values to true values and assessing predictive accuracy, i.e. focussing on the wrong thing? I agree that it can be a useful way to understand why a method performed poorly but don’t think that means it is always necessary (again consider simple mean imputation of an incomplete covariate in trials).

* **We have adopted the suggestion.**

Section 3.6 suggests three performance measures. I was glad to see this, though my own ‘default’ set is slightly different. The first question is, why these three? The second is whether these are the authors’ preferences for all simulation studies or if there is anything specific for simulation studies of imputation techniques? Personally I don’t think the choice of performance measures should have anything to do with the methods being evaluated but the aims of the analysis in practice.

* **We have emphasized that the evaluations should match the missing data problem under investigation.**

A few comments on table 5:

a) ‘Comparative truth’ seems an inappropriate heading here and makes it hard to put ‘induced missingness’, part of the data-generating mechanism, with ‘data-generating mechanism’. Comparative truth is really relevant to the estimands/targets of interest.

b) Would be good to include ‘full-data’ analysis under ‘reference method’.

🡪 **The revised manuscript does not longer include this table. TODO: check Table 1!**

MISC. & MINOR COMMENTS

There are two points that arise repeatedly in my collaborations but that are not currently discussed in the paper:

1. ‘Accidental MNAR’. People intend to simulate MAR but do something like missingness in X2 depends on values of X1 and missingness in X3 depends on X1 and full-data values of X2 (whether or not X2 is observed). Because X2 may be missing, X3 is MNAR. Personally I think this is common enough to comment on and is a particular issue for non-monotone missing data patterns (links to Robins and Gill’s work on randomised monotone missingness).

2. Fixing the proportion of missingness. This is often done to avoid large variation in the proportion missing across repetitions, which is understandable. In doing so, missingness is no longer independent across observations (by observation i being missing the probability of other observations i– being missing is reduced) but I have never seen this cause problems.

**🡪 TODO: maybe include this still??**

Minor point: The description of congeniality in section 2.3 is incorrect. Meng’s definition is not about getting the true non-response mechanism right. It’s about being able to embed the analysis procedure and the imputation model within a Bayesian model. So we may misspecify the nonresponse mechanism but have a congenial multiple imputation procedure.

* **We have changed the description in line with the suggestion.**

Section 3.6 – This is a really picky comment but technically \*estimates\* do not have any properties, like unbiasedness. So it’s a MI estimator/method that we want to be unbiased. And it’s the confidence interval (not the estimates) that we want to have the nominal coverage proportion.

* **We have adopted the suggestions.**

‘golden standard’ should say ‘gold standard’

* **Thanks!**