This is a nicely written paper giving advice on how to design simulation studies for a standardized evaluation of imputation methodology. It is non-technical and useful. I do however have concerns regarding some arguments made:

Major:

- I think especially the arguments of Section 2 confuse two distinct issues: problems generated by a method and problems generated by inadequate use of the methods from a human. Similarly, there is a confusion about the problems generated by methods and the problems generated by inadequate generalization of humans. I give some examples: Section 2.1.: whether simulation results are generalizable from a given population/data-generating process (DGP) is a data fusion/generalizability/transportability issue but not an issue on how data a generated in a simulation study. Section 2.2: whether method A (e.g. imputation) outperforms method B (e.g. listwise deletion) depends (as later stated by the authors) on the estimand, purpose of the study, used distributions (in the DGP, imputation model, possibly also analysis model), and simulation setup (e.g. dependency structure) – if a human misinterprets findings in a comparison of method A and B, then this is a cognitive problem and not the problem on how missingness was generated in the DGP (some of those reflections can be found in the papers by Sander Greenland). I recommend reflecting this dimension particularly in Section 2 and make clearer distinctions on what actually causes problems, which includes interpretation and generalization of selected simulation studies.

Minor:

- Table 1. I understand the paper is non-technical, but I would give a more precise definition of MCAR/MAR/MNAR or at least give good references. I recommend for example Doretti et al. (1), and also Seaman et al. and Mealli et al. (2, 3)
- Section 2.3., second last paragraph and Section 3.6 (ii): the authors recommend that confidence intervals should have greater or equal 95% coverage (as it is often recommended). Ideally, the aim should be that actual coverage equals nominal coverage, for example 95%. I would argue that in most cases it would be preferable to have 94% coverage than 100% - and this is line with the authors argument of CI width. Maybe a more refined advice could be given (optional)?
- Section 3.2: "as a general rule it would be wise to include one or more multivariate estimands". I disagree. If there is a clear scientific question or problem, then one has a clear and unambiguous estimand (4), and if this is say a marginal (causal) risk ratio, then there is no need to evaluate say, for example, a conditional odds ratio or a completely unrelated estimand. If a simulation study is inspired by a given question, then it should focus on the estimand of interest.
- Section 3.3.: I find the simple MNAR – MAR – MCAR distinction not ideal. While certainly not the main aim of the paper, it should at least be outlined that this distinctions has several problems, including the verification and motivation of MAR when multiple variables are missing and the fact that for given DGP's a complete case analysis may be valid under MNAR but not MAR. The best reference for those arguments is the Mohan and Pearl paper (5) (see also other related references (6, 7)). Briefly, readers should be at least aware that this distinction can have problems and where to read more. Of course, the authors are right that "missingness should be induced according to several sets of missing data conditions"
- Table 4: while I like rules of thumb, I am unsure whether with 50% missingness we necessarily get bad results if n is large. Can the authors give references for the statements of Table 4?

- Section 3.5 (ii): In Amelia II this is actually implemented and described in the accompanying paper could be referenced. (8)
- Section 3.5, (iv): Can you make a stronger argument? I am not convinced it is a good one. The idea of imputation is to make valid inference about some parameter of interest, and while there is no harm in it, I can`t see why imputed values need to be plausible beyond meeting basic requirements (say respecting bounds).
- Section 3.6.: I would add item (iv) on prediction and performance evaluation through measures such as the mean squared prediction error, and others. If the aim is purely predictive, and data are missing, I find it convincing to simply evaluate (out of sample) prediction errors.

1. Doretti M, Geneletti S, Stanghellini E. Missing Data: A Unified Taxonomy Guided by Conditional Independence. International Statistical Review. 2018;86(2):189-204.
2. Mealli F, Rubin DB. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika. 2015;102(4):995-1000.
3. Seaman S, Galati J, Jackson D, Carlin J. What Is Meant by "Missing at Random"? Stat Sci. 2013;28(2):257-68.
4. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014;25(3):418-26.
5. Mohan K, Pearl J. Graphical Models for Processing Missing Data. J Am Stat Assoc. 2020;in press.
6. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(12):2705-15.
7. Schomaker M. Regression and Causality. arXiv e-prints. 2020;https://arxiv.org/abs/2006.11754.
8. Honaker J, King G, Blackwell M. Amelia II: A Program for Missing Data. J Stat Softw. 2011;45(7):1-47.