PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine
AUTHORS	Waljee, Akbar; Mukherjee, Ashin; Singal, Amit; Zhang, Yiwei; Warren, Jeffrey; Balis, Ulysses; Marrero, Jorge; Zhu, Ji; Higgins, Peter

VERSION 1 - REVIEW

REVIEWER	Angela Wood Lecturer in Biostatistics University of Cambridge UK I have no competing interests.
REVIEW RETURNED	27-Mar-2013

THE STUDY	Study design Two studies are used to compare imputation methods for missing laboratory data. Both studies are analysed using a random forest model, which is likely to favour the missForest imputation approach over the other imputation approaches. In addition, I disagree that “an extensive simulation study” has been performed. The results are based on averages of only 30 datasets, which differ by random selections of missing values. An extensive simulation study would typically include more random variations in the datasets, for example, using bootstrap samples or simulated outcomes and use 1000 simulated datasets to reduce Monte Carlo errors. Representative of evidence The conclusions of the research are not generalisiable to all studies/analyses with missing laboratory data. Methods description Statistical terminology has not been used well. For example, in the article summary, MICE is defined as “multiple imputation by chained equations” but in the manuscript it is defined as “multivariate imputation by chained equations”. Due to a lack of detail, I expect multiple imputation has not been used, and thus the authors should be very clear in the abstract and throughout they are applying a SINGLE regression imputation approach using MICE. Terminology for “missing at random” and “missing completely at random” is also confused in the manuscript. I expect the authors have simulated “missing completely at random” values in their datasets. This needs to be clarified throughout.

	Abstract/summary The abstract is written to suggest that the findings are generalisable to all studies, whereas the summary is written with more reference to the two specific studies. The latter is more appropriate. The limitations and strengths should focus on the limitations and strengths of the simulation study and not on missForest. Are the statistical methods appropriate? Given that the outcome models of both datasets are random forest models, it is not surprising that missForest out performs the other imputation approaches. It is important that imputation and analysis models have some degree of congeniality. Imputations methods such as MICE, which make parametric assumptions about the relationships between observed variables, are not congenial with random forest models. This needs discussion. In the description on page 8 it seems that the authors have used "multivariate imputation by chained equations" – and thus multiple imputation has not been applied. I expect a single imputation has been created from the package "mice". This is a limited approach and not one that is commonly used. Usually, one would create multiple imputations from the MICE approach. I recommend the authors repeat their simulations using multiple imputations for MICE.
RESULTS & CONCLUSIONS	The four different imputation methods have been compared in two small datasets but the findings cannot be generalised to all missing laboratory data problems. Due to the fact that the two small datasets are likely to favour the missForest method, the results are not credible and the results need more careful interpretation. Thus, the overall message of the manuscript is misleading.

REVIEWER	Brian Jackson Associate Professor of Pathology (Clinical), University of Utah, USA I have no competing interests to declare.
REVIEW RETURNED	27-Mar-2013

GENERAL COMMENTS	Nice article on a practical topic of emerging interest to the diagnostics community. My only thoughts: 1. Ideally this article should be reviewed by someone with expertise in the methodology discussed in the article. I have a math background, but don't have specific experience with these particular algorithms and I'm not sure I would recognize any possible limitations that that authors could have omitted. 2. It seems to me that a key issue here is generalizability of the results to other MAAAs. I think the authors should clarify in the abstract and article summary that this study measured the accuracy of imputation methods based on two specific MAAAs, rather than saying that it measured accuracy in general. Then in the Discussion section it would be reasonable to lay out the argument for why these findings might be more broadly generalizable. Also, I would rework the figures to show the HCC study findings alongside the thiopurine study findings right in the same figures. This would give a better visual sense of how sensitive the findings are to the choice of MAAA.

REVIEWER	Jason M Baron, MD Fellow, Pathology Informatics Department of Pathology Massachusetts General Hospital I have no competing interests to declare.
REVIEW RETURNED	28-Mar-2013

GENERAL COMMENTS	Mukherjee et al. address an important and sometimes underappreciated challenge in predicting clinical outcomes from medical data using machine learning techniques: applying prediction models in the setting of missing predictor data. In particular, they address the robustness of two specific multi-analyte assays with algorithmic analyses (MAAAs), when missing data are imputed, and compare four different previously available and implemented imputation techniques. Starting with complete sets of test data (no missing data) from real patients, they simulate missing data by redacting randomly selected data points. They evaluate the accuracy of MAAA predicted clinical outcomes when the redacted data is imputed using each of the four different imputation techniques, taking model performance with the original (complete) dataset as the goldstandard. They find that the "missForest" method of imputing data generally outperforms the other imputation methods tested both in terms of imputation accuracy and the accuracy of the MAAA clinical output. Overall, their methods are sound and their results seem reasonable. Their findings are interesting and useful and are deserving of publication. However, the authors may consider several minor revisions to improve the manuscript and its conclusions: 1) As the authors acknowledge in the introduction, missing data are rarely truly random. Moreover, real data could contain associations between which values are missing and the clinical outcome of interest or the specific values for certain predictors. For example, a clinician might be more likely to order a test when an abnormal result is expected based upon clinical circumstance. Non-randomness of missing data has the potential to lead to imputation bias. To address these issues, the authors may consider strengthening their findings and the manuscript by doing one of the following: i) They could test the imputation techniques when non-randomly missing data is simulated. Ideally, this analysis would use empirically derived information about the association between predictors or outcomes and the likelihood that particular elements data will be missing. Comparing model performance when non-random missing data is simulated and imputed to performance when random noise is added to predictors may also help to tease out the specific effects of imputation bias on MAAA model accuracy. ii) Alternatively, they could include this issue of non-random missing data in greater detail in the discussion section, noting it as a potential limitation. As appropriate and feasible, they could consider speculating on how non-random missing data might impact their results and whether the four different techniques would perform similarly with non-random missing data. While doing point i) might provide very useful data and strengthen the manuscript, I would consider this optional and NOT required for publication.

	2) The authors may wish to revise the scale of the Y-axis in the first and third figures to start at 0. As they are now, these figures may be slightly misleading in exaggerating the differences. 3) The authors should consider adding P-values to figures to denote the significance of differences in accuracy between the four different imputation techniques. 4) In the second paragraph of the results section, the authors presumably mean “categorical” instead of “continuous”, the second time the word "continuous" is used. 5) The authors mention an increase in computation time using missForest in the range of 10-20 seconds. Since processing times are dependent on the particular computing platform, they should either note the specific platform used or remove the references to specific lengths of time. They may also considering mentioning how processing time would be expected to scale for larger datasets if this is of practical relevance.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1 (Angela Woods):

1. Two studies are used to compare imputation methods for missing laboratory data. Both studies are analysed using a random forest model, which is likely to favour the missForest imputation approach over the other imputation approaches.

This will be addressed in point 7 below.

2. In addition, I disagree that “an extensive simulation study” has been performed. The results are based on averages of only 30 datasets, which differ by random selections of missing values. An extensive simulation study would typically include more random variations in the datasets, for example, using bootstrap samples or simulated outcomes and use 1000 simulated datasets to reduce Monte Carlo errors.

As suggested by the reviewer a more extensive simulation was performed that involved the use of a 1000 datasets. The results are based on the average of these 1000 datasets. New figures have been generated that also reflect both continuous and categorical variables.

3. Statistical terminology has not been used well. For example, in the article summary, MICE is defined as “multiple imputation by chained equations” but in the manuscript it is defined as “multivariate imputation by chained equations”. Due to a lack of detail, I expect multiple imputation has not been used, and thus the authors should be very clear in the abstract and throughout they are applying a SINGLE regression imputation approach using MICE.

We used multivariate imputation by chained equations (MICE) and we have made those corrections in the manuscript.

4. Terminology for “missing at random” and “missing completely at random” is also confused in the manuscript. I expect the authors have simulated “missing completely at random” values in their datasets. This needs to be clarified throughout.

The terminology has been clarified in the entire manuscript and the term “missing completely at

random" has been used throughout the manuscript.

5. The abstract is written to suggest that the findings are generalisable to all studies, whereas the summary is written with more reference to the two specific studies. The latter is more appropriate.

The abstract has been revised to reflect that this method has been tested in two specific studies.

6. The limitations and strengths should focus on the limitations and strengths of the simulation study and not on missForest.

The following sentences have been added to the discussion section to reflect the limitations and strengths of the simulation study.

"We have performed an extensive simulation study using two clinical datasets and two published predictive models to compare the performance of four methods of missing value imputation for missing data completely at random. We included both local (randomForest) and global (logistic) modeling approaches to avoid bias that might favor a local (MissForest) imputation approach. While the superiority of MissForest for imputation of missing lab values this will not be generalizable to all predictive models or datasets, this manuscript highlights the value of missForest to impute missing data. We compared four popular methods namely, missForest, Nearest Neighbor, MICE and mean imputation, in two studies simulating data missing completely at random. We found that these simulation methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values. In addition, the ready availability of the freeware R package makes missForest and its simulations a very convenient solution for any practical missing value problems. The main limitations of these simulations as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods. An additional limitation in this study is that these simulations did not address the issue of data missing for non-random reasons. There could be an association between the clinical outcome of interest and the missingness of certain predictors. At this point, we cannot generalize these results to situation in which data is missing for non-random reasons."

7. Given that the outcome models of both datasets are random forest models, it is not surprising that missForest out performs the other imputation approaches. It is important that imputation and analysis models have some degree of congeniality. Imputations methods such as MICE, which make parametric assumptions about the relationships between observed variables, are not congenial with random forest models. This needs discussion.

We appreciate the reviewer's comments. Random forest approaches are a local model where as logistic statistical techniques are a more global model. missForest might be better for a random forest model because it is local; however, we showed it is also superior when used with a global model such as logistic regression. missForest overall is a more powerful and flexible imputation method. In order to highlight this we incorporate both the logistic and random forest approaches in the statistical analysis and results section.

8. In the description on page 8 it seems that the authors have used "multivariate imputation by chained equations" – and thus multiple imputation has not been applied. I expect a single imputation has been created from the package "mice". This is a limited approach and not one that is commonly used. Usually, one would create multiple imputations from the MICE approach. I recommend the authors repeat their simulations using multiple imputations for MICE.

This has been addressed in point 3 above.

9. The four different imputation methods have been compared in two small datasets but the findings cannot be generalised to all missing laboratory data problems. Due to the fact that the two small datasets are likely to favour the missForest method, the results are not credible and the results need more careful interpretation. Thus, the overall message of the manuscript is misleading.

The abstract and the discussion has been edited to reflect that this methodology works in the two clinical examples that we used and that it may be an option for other clinical prediction models.

The reviewer is correct in that we simulated missing completely at random values in the dataset. This has been clarified in the manuscript.

Reviewer #2 (Brian Jackson):

1. Ideally this article should be reviewed by someone with expertise in the methodology discussed in the article. I have a math background, but don't have specific experience with these particular algorithms and I'm not sure I would recognize any possible limitations that that authors could have omitted.

No Revisions Needed

2. It seems to me that a key issue here is generalizability of the results to other MAAAs. I think the authors should clarify in the abstract and article summary that this study measured the accuracy of imputation methods based on two specific MAAAs, rather than saying that it measured accuracy in general. Then in the Discussion section it would be reasonable to lay out the argument for why these findings might be more broadly generalizable.

The aims in the article summary, abstract and manuscript reflect that the accuracy of the four different imputation methods is based on two large datasets and that the accuracy of the prediction is based on these datasets. In addition, the following paragraph has been added to the discussion section:

“With increasing complexity of these models, and increasing numbers of analytes, the risk of missing completely at random values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs and may be applicable to other datasets with missing completely at random datasets. “

3. Also, I would rework the figures to show the HCC study findings alongside the thiopurine study findings right in the same figures. This would give a better visual sense of how sensitive the findings are to the choice of MAAA.

The figures have been revised to reflect a comparison of the logistic regression and the random forest. Due to limitation in the number of figures we opted to highlight the differences in the logistic regression compared to random forest rather than the different models based on the comments from all the reviewers. We would be more than happy to show those comparisons in the supplemental section as a panel of 8 figures if the reviewers prefer that too.

Reviewer #3 (Jason Baron):

1. As the authors acknowledge in the introduction, missing data are rarely truly random. Moreover, real data could contain associations between which values are missing and the clinical outcome of interest or the specific values for certain predictors. For example, a clinician might be more likely to order a test when an abnormal result is expected based upon clinical circumstance. Non-randomness of missing data has the potential to lead to imputation bias.

To address these issues, the authors may consider strengthening their findings and the manuscript by doing one of the following:

i) They could test the imputation techniques when non-randomly missing data is simulated. Ideally, this analysis would use empirically derived information about the association between predictors or outcomes and the likelihood that particular elements data will be missing. Comparing model performance when non-random missing data is simulated and imputed to performance when random noise is added to predictors may also help to tease out the specific effects of imputation bias on MAAA model accuracy.

ii) Alternatively, they could include this issue of non-random missing data in greater detail in the discussion section, noting it as a potential limitation. As appropriate and feasible, they could consider speculating on how non-random missing data might impact their results and whether the four different techniques would perform similarly with non-random missing data.

While doing point i) might provide very useful data and strengthen the manuscript, I would consider this optional and NOT required for publication.

We appreciate these comments and the importance of addressing the non-randomness of missing data. We have added the following paragraph to the discussion section as suggested by the reviewer.

“An additional limitation in this study is the concern regarding the non-randomness of missing data and the concern that there may be an association between the clinical outcome of interest and the specific values of certain predictors. In these illustrations, where the emphasis is on missing completely at random lab data, the labs are drawn as a group of individual labs and where one lab would be missing due to an error in processing rather than a complete set of missing lab data.”

2. The authors may wish to revise the scale of the Y-axis in the first and third figures to start at 0. As they are now, these figures may be slightly misleading in exaggerating the differences.

The figures have been revised and now start at 0

3. The authors should consider adding P-values to figures to denote the significance of differences in accuracy between the four different imputation techniques.

A comment has been placed in the figures to denote that:

“*missForest vs. all other models have $p<0.0001$”

4. In the second paragraph of the results section, the authors presumably mean “categorical” instead of “continuous”, the second time the word "continuous" is used.

The word categorical has been substituted in reference to the appropriate graph.

5. The authors mention an increase in computation time using missForest in the range of 10-20 seconds. Since processing times are dependent on the particular computing platform, they should either note the specific platform used or remove the references to specific lengths of time. They may also considering mentioning how processing time would be expected to scale for larger datasets if this is of practical relevance.

The reference to the computation time has been removed.

VERSION 2 – REVIEW

REVIEWER	Angela Wood Lecturer in Biostatistics University of Cambridge UK I have no competing interests.
REVIEW RETURNED	04-Jun-2013

THE STUDY	Statistical Methods I would personally prefer to see a little more detail regarding the imputation techniques (eg, what variables were included in the imputation models). There is a slight lack of transparency regarding this. However, I appreciate this may not be of interest to most readers, and the authors have appropriately referenced the techniques. I think the authors should be clear that MICE has been used to perform a single imputation. The authors title the paper "Modern imputation techniques". I suggest removing "Modern". MICE is rarely used for single imputation and mean imputation is not modern.
GENERAL COMMENTS	The authors have adeqautely addressed my previous concerns. My final comment concerns the suitability of this statistical methods comparison study for publication in BMJ open. I would suggest it is more suitable for publication in a statistical journal.

REVIEWER	Jason Baron, MD Fellow, Pathology Informatics Department of Pathology Massachusetts General Hospital
REVIEW RETURNED	31-May-2013

GENERAL COMMENTS	This manuscript offers an important contribution to the field. This revised version seems suitable for publication.