Associate Editor

Comments to the Author:

Both reviewers feel that this paper needs a lot of further explanation, and even modifications to the underlying package. If you feel that you can satisfy all of their concerns, please include a detailed response with your revision.

Dear Dr. Sarkar,

We would like to thank you for your response and the invitation to resubmit a revised version of our manuscript "smdi: An R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies," submitted to JAMIA Open (Manuscript ID: JAMIO-2023-0165).

We are thankful for the constructive and thoughtful comments provided by the reviewers in evaluating our work. We acknowledge the valuable insights, have refined the manuscript and provided updates to the underlying package to align with the reviewers' recommendations. Please see our detailed responses one by one below.

We believe that these revisions have significantly enhanced the quality and contribution of our work which we feel is now well-suited for publication in JAMIA Open.

Thank you for your consideration of our revised manuscript.

Sincerely,

Janick Weberpals, RPh, PhD

# Reviewer: 1

Comments to the Author

This application notes paper addresses an important topic – the impact of actual and perceived missingness in EHR data. The authors apply statistical methods in the design of a new R package that analyzes data for patterns of random or non-random missingness.

The manuscript focuses exclusively on the design and execution of the R package. The results are descriptive in the “we developed this” vein than a description of the findings from running smdi. The tables and figures show they have a limited body of actual results that could be addressed more directly in the results section and the discussion sections.

[RESPONSE] We thank the reviewer for the valuable feedback on our manuscript. We acknowledge the need to more deeply discuss the results obtained through the utilization of the respective smdi functions. To address the reviewer’s comment, we added a paragraph to each results section discussing the outputs, tables, and figures more directly and further added a brief discussion to aid readers with the interpretation of the results.

**Section Inferential Three Group Diagnostics, Group 1 Diagnostics (last paragraph):** *“Applying group 1 diagnostics to the synthetic example dataset would indicate that the ECOG POC (median ASMD 0.03, min-max 0.00-0.07, p-value 0.78) does not show any differences in observed patient characteristics between patients with and without and observed value for ECOG which would give evidence for a MCAR mechanism (Figure 3 bottom, Group 1 diagnostics - orange boxes) . Conversely, in the case of EGFR and PD-L1, the group 1 diagnostics display larger differences and hence may rather underlie a MAR or MNAR mechanism (Figure 2 c, Figure 3).”*

**Section Inferential Three Group Diagnostics, Group 2 Diagnostics (last paragraph):** *“Figure 3 (Group 2 diagnostics - blue boxes), for example, illustrates the AUC values of the output of smdi\_rf() when applied to the synthetic example dataset. Since the missingness of the EGFR POC follows a true MAR mechanism, the resulting AUC of 0.63 is expectedly meaningfully higher than what is observed for ECOG (0.51) and PD-L1 (0.52) which follow a true MCAR and MNAR mechanism, respectively.”*

**Section Inferential Three Group Diagnostics, Group 3 Diagnostics (last paragraph):** *“Demonstrating the utilization of smdi\_outcome() using the example dataset, the derived logHR coefficients for the missingness indicators of the POCs (Figure 3 bottom, Group 3 diagnostics - green boxes) align with the anticipated outcomes from our simulations [22]. Specifically, EGFR manifests no discernible difference in the outcome after adjustment for fully observed covariates (logHR -0.01, 95% confidence intervals [CI] -0.10 to 0.09), suggesting a MAR mechanism. ECOG exhibits no distinction in either the unadjusted or adjusted model (logHR -0.06, -0.16 to 0.03), indicating MCAR. Conversely, PD-L1 showcases differences in the outcome in both models, suggesting an MNAR context.”*

A significant concern is that the smdi package, as described, doesn’t have logic that reflects the nuances of EHR data (and the workflows that generate the data).

[RESPONSE] We thank the reviewer for this comment and understand the importance of reflecting the nuances specific to EHR data in our package's logic. During the design and validation of the package functions, we employed comprehensive simulations and a real-world example (corresponding manuscript is currently under review) following assumptions on missingness generating mechanisms that we feel are realistic and expected when working with EHR data. To provide transparency on these assumptions and to reflect this aspect in the manuscript, we added these details to the “**Supplementary Methods**” section (**sub-chapter 1.1. Missingness assumptions**).

However, we agree that these mechanistic assumptions and the derived smdi diagnostics always need to go hand-in-hand with domain expertise knowledge coming from stakeholders involved in the data generating process such as clinicians, nurses, pharmacists and others. To reflect this aspect, we enhanced the discussion highlighting this aspect in the “Conclusions” section of the manuscript:  
  
**Discussion (last paragraph):** *“The package should be used with certain limitations in mind. Most importantly, the true underlying mechanism causing the missing data can never be inferred with absolute certainty from the observed data. Therefore, it is important to complement diagnostic results with substantive expert knowledge to factor in how covariates are measured in routine care, which could be system-specific, and contextualize potential reasons for missingness. This collaborative approach allows for a contextualized understanding of potential causes for missing data in EHR.”*

A major gap is the reference to a claims data set without specifying the data source.

[RESPONSE] We thank the reviewer for pointing this out. We added more details and references on the databases that were initially used to perform the simulation study that served as the foundation for the design of the smdi R package.

**Background and Significance (last paragraph):** *[..] “Considering these limitations, we have recently developed and evaluated a principled approach combining multiple missing data diagnostics [22] using a database linkage from the Mass General Brigham Research Patient Data Registry EHR in Boston [23] linked with Medicare fee-for-service claims data [24].”*

*22. Weberpals J, Raman SR, Shaw PA, et al. A principled approach to characterize and analyze partially observed confounder data from electronic health records. Submitted. 2023.*

*23. Nalichowski R, Keogh D, Chueh HC, et al. Calculating the benefits of a research patient data repository. AMIA Annual Symposium Proceedings. 2006;2006:1044.*

*24. https://resdac.org/ (accessed 16 November 2023)*

The authors provide a very detailed list of disclosures and reference FDA approval but the reasons for any of these are not apparent in the body of the manuscript.

[RESPONSE] We thank the reviewer for this comment. As disclosed in the manuscript document, this project was supported by Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA) and the development of this R package has been a part of this effort. This project is additionally part of a larger initiative to integrate electronic health record data in a causal analysis framework as described in more detail by Desai et al.1

To provide more context in the body of the manuscript we rephrased the introduction and added this aspect.

**Background and Significance (first paragraph):** *“Administrative health insurance claims databases have traditionally been the foundation for pharmacoepidemiologic studies. However, a notable drawback lies in their inability to capture detailed clinical nuances. To overcome this limitation, substantial initiatives are underway, for instance in the FDA Sentinel initiative1,linking claims databases and electronic health records (EHR) to generate real-world evidence (RWE) and complement data from randomized controlled trials (RCTs).1,2*

*1.* Desai RJ, Matheny ME, Johnson K, et al. Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework. Npj Digit Med. 2021;4(1):170. doi:10.1038/s41746-021-00542-0

2. United States Food and Drug Administration. Framework for FDA’s real world evidence program. Dec 2018. Accessed 6/30/2023.

# Reviewer: 2

Comments to the Author

- 1. This article introduces a toolkit designed to be a valuable addition to routine healthcare database analysis, focusing on the study of structural missing data. The toolkit enables users to systematically investigate missing data on partially observed confounders by providing functions for visualizing, describing, and inferring potential mechanisms based on observed data. It also offers a synthetic dataset, comprehensive documentation, examples, and support to assist users in effectively utilizing the toolkit. This toolkit holds the potential for practical applications and contributions to healthcare database analysis.

[RESPONSE] We thank the reviewer for recognizing our work and for the constructive feedback on our manuscript.

- 2. In the 'Materials and Methods' section of the paper, the authors need to include explanations for key statistical principles and methods employed in the toolkit's development. These explanations should cover their relevance to the research and any specific formulas or calculations involved. It is important to discuss the advantages of using these methods in the context of the toolkit's objectives. This additional content will help readers better understand the toolkit's scientific foundation. (Major)

[RESPONSE] We thank the reviewer for pointing this out and acknowledge the need to provide further details on the statistical principles and methods that underlie the design of the smdi R package. To address this concern, we compiled a comprehensive **Supplementary Material** document and added details in **sub-chapters “1.1 Missingness assumptions”** and “**1.2 Missingness characterization”** where we illustrate all real-world missingness assumptions using directed acyclic graphs and outline key statistical principles including formulas, references and further details on the package functions. We refer to this **Supplementary Material** in the main manuscript text in the last paragraph in the “Results > Main Package Functions” sub-chapter:

**Results, section Main package Functions (last paragraph):** “*Details on missingness assumptions, key statistical principles and further information on all functions can be found in the Supplementary Methods and in the documentation of each respective function which can be accessed in R by preceding the function name with a question mark, e.g.: ?smdi\_diagnose()”*

- 3. For the smdi\_rf(): Random Forest Classification Model, if the toolkit utilizes a random forest classification model, it is essential to explain the rationale behind its selection, provide details on the training process, and specify the model parameters. Additionally, insights into model performance assessment, including cross-validation techniques, should be provided.

[RESPONSE] We thank the reviewer for highlighting the lack of details regarding the implementation of the random forest model to diagnose missing data. We opted for a random forest model as this type of machine learning model has many beneficial features, including:

* Ability to implicitly model nonlinear and non-additive relationships between observed variables (i.e., higher order terms do not need to be explicitly specified).
* Recursive partitioning models like random forests have been found to work particularly well with sparse tabular data, which is the typical data type that is used for real-world evidence studies.
* Random Forests provide easy-to-interpret information on feature importance, indicating the variables contributing significantly to predicting missingness. This aids in identifying key features driving the missingness patterns in the dataset.
* Multiple other studies have also reported promising results of this random forest-based approach.1,2

We initially kept the hyperparameter tuning in smdi\_rf rather simple as diagnosing missing data primarily involves the identification of patterns/mechanisms and relationships related to missingness, rather than maximizing prediction accuracy. Therefore, the fine-tuning of hyperparameters for optimal predictive performance was not the primary objective of this function. But we agree with the reviewer that it would be beneficial to give users the flexibility to tune hyperparameters if they wish to do so. We hence implemented an additional “tune” parameter which, if set to tune = TRUE, will perform a 5-fold cross validation and a random search for the optimal number of variables randomly sampled as candidates at each split (mtry). We also updated the function to report the estimated OOB error for each investigated partially observed confounder.

These updates are reflected in the [developer “dev” branch](https://gitlab.partners.org/janickweberpals/smdi/-/tree/dev?ref_type=heads) , which can be installed via

devtools::install\_git("https://gitlab-scm.partners.org/janickweberpals/smdi.git", ref = "dev")

and will be submitted to CRAN with the next release cycle (version 0.3.0) after acceptance of this manuscript.

As suggested by the reviewer, we now provide more technical details on the implementation of the random forest model in the smdi package and discuss our reasons to implement random forest models for the Group 2 diagnostics in the **Supplementary Methods sub-chapter “1.2 Missingness characterization** > smdi\_rf()”.

1 Sondhi A, Weberpals J, Yerram P,et al. A systematic approach towards missing lab data in electronic health records: A case study in non-small cell lung cancer and multiple myeloma. CPT: Pharmacometrics & Systems Pharmacology. Published Online First:15 June 2023.

2 Beaulieu-Jones BK, Lavage DR, Snyder JW, et al. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. JMIR Medical Informatics. 2018;6:e11.7

- 4. smdi\_outcome() :In addition to linear regression, logistic rgression, and the Cox proportional hazards model, is it possible to consider supporting other types of regression models, such as variants of Generalized Linear Models (GLM), additional models in survival analysis, and even deep learning models. This would increase the versatility of the tool, enabling it to handle a wider range of types of outcome data.

[RESPONSE] We thank the reviewer for this comment and agree it would be a useful feature to add additional options of potential GLM outcome models to this function. To meet the reviewer’s comment, we have deprecated the model = “logistic” argument in lieu of model = “glm” and added an additional parameter *glm\_family* to specify eligible family objects of *glm* models as defined by stats::family() and supported by stats::glm(). Eligible family objects are:

* binomial(link = "logit")
* gaussian(link = "identity")
* Gamma(link = "inverse")
* inverse.gaussian(link = "1/mu^2")
* poisson(link = "log")
* quasi(link = "identity", variance = "constant")
* quasibinomial(link = "logit")
* quasipoisson(link = "log")

Deep learning models are not very common in real-world evidence studies and we feel the implementation, dependency management and maintenance of such complex models would create an overhead that may not be justified for the utility they would provide in context of this package. Moreover, more complex neural network-based models have only limited advantages in sparse tabular data and a growing body of literature suggests that recursive partitioning models or regression models perform just as well or better in most of the use cases relevant for real-world evidence studies.1-4

The above described updates to *smdi\_outcome()* are already reflected in the [developer “dev” branch](https://gitlab.partners.org/janickweberpals/smdi/-/tree/dev?ref_type=heads) , which can be installed via

devtools::install\_git("https://gitlab-scm.partners.org/janickweberpals/smdi.git", ref = "dev")

and will be submitted to CRAN with the next release cycle (version 0.3.0) after acceptance of this manuscript.

We modified the corresponding section in the manuscript accordingly.

**Section Inferential Three Group Diagnostics, Group 3 Diagnostics (second to last paragraph):** *“smdi\_outcome() supports multiple outcome regression types: including linear regression (lm [37]) for continuous outcomes, Cox proportional hazards model (coxph [35]) for time-to-event outcomes, and generalized linear regression models (glm [37]) for which the family of conditional distributions of the outcome can be selected using the glm\_family parameter (the default is binomial(link='logit')). Besides the regression type (model parameter) and the glm\_family (in case of a glm model), users need to specify the column containing the outcome using the form\_lhs parameter (e.g., Surv(eventtime, status) in case of a Cox model).”*

1 Weberpals J, Becker T, Davies J, Schmich F, Rüttinger D, Theis FJ, Bauer-Mehren A. Deep Learning-based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-scale, Real-world Data Study. Epidemiology. 2021 May 1;32(3):378-88.

2 Loureiro H, Becker T, Bauer-Mehren A, Ahmidi N, Weberpals J. Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study. Front Artif Intell. 2021 Apr 16;4:625573.

3 Shwartz-Ziv R, Armon A.Tabular data: Deep learning is not all you need. Information Fusion. 2022;81:84–90.17

4 Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? 2022.https://arxiv.org/abs/2207.0881518

- 5. The paper mentions three functions: smdi\_check\_covar(), smdi\_na\_indicator(), and md\_pattern(), as part of the toolkit. However, it lacks detailed explanations. I suggest that the authors include information about the specific functionality and differences between these functions within the paper.

[RESPONSE] We thank the reviewer for making us aware of this. *smdi\_check\_covar(), smdi\_na\_indicator()* are utility functions which are called in the background by other smdi functions but which can also be used as standalone functions if desired by the user. We now provide more details on these functions in the **Supplementary Methods** sub-section “Utility functions”.

The *md.pattern()* function, a re-export of the mice package, is similar to *gg\_miss\_upset()* and returns a matrix displaying the frequency of each observed missing data pattern. We added a short description to the main manuscript text and refer for more details to the **Supplementary Methods** section where we provide further information.

**Descriptives and Pattern Diagnostics (last paragraph):** *“The md.pattern() function, a re-export of the mice package [29], provides a similar functionality and returns a matrix displaying the frequency of each observed missing data pattern.”*

- 6. To diagnose missing mechanisms based on observed data, researchers need to integrate Group 1 diagnostics, Group 2 diagnostics, and Group 3 diagnostics. This process can be somewhat inconvenient. Creating an additional function that can directly provide the diagnosis result (MCAR, MAR, MNAR, etc.) would greatly enhance convenience for researchers.

[RESPONSE] We thank the reviewer for pointing this out. We agree with the reviewer and would like to point out sub-section “smdi\_diagnose() to compute all three group diagnostics” in the manuscript. The *smdi\_diagnose()* function is a wrapper around group diagnostics and directly provides results on all the three inferential group diagnostics with a single function call. In addition, we added a short sub-paragraph detailing how the ancillary *smdi\_style\_gt()* function can be used to automatically format and export the table that is being returned by *smdi\_diagnose()* as a publication-ready table*.* This enables users to practically create, format and export the results of all three group diagnostics within two lines of code. However, as with any statistical hypothesis testing, we refrain from adding a function that hardcodes the interpretation of the diagnostic results as this is not always straightforward and should be done in the context of the specific healthcare system. As pointed out in the second comment of reviewer #1 (and discussed in the last paragraph of the discussion section), the diagnostics are based on mechanistic assumptions and should ideally go along with domain expertise knowledge coming from stakeholders involved in the data generating process such as clinicians, nurses, pharmacists and others. Based on the first comment of reviewer #1, we now provide examples of possible interpretations of the diagnostic outputs throughout the manuscript’s results section.

**Inferential Three Group Diagnostics,** sub-section **smdi\_diagnose() to compute all three group diagnostics:**

*“Finally, the smdi\_diagnose() function enables users to compute all of the above discussed group diagnostics within a single function call.*

*# minimal example of a smdi\_diagnose() function call*

*smdi\_diagnose(*

*data = smdi\_data,*

*model = "cox",*

*form\_lhs = "Surv(eventtime, status)",*

*n\_cores = 3*

*)*

*The function returns an object of class smdi containing a table with the results of all diagnostics for each specified POC and Little’s test p-value across all covariates (Table 2). By cross-checking all resulting diagnostic parameters to expected estimates as illustrated in in the above examples (Figure 3, [22]), the diagnostics can provide valuable insights into underlying missingness mechanisms and thereby help elucidate if analytic approaches such as imputation analyses are viable options.*

*The smdi\_style\_gt() function is an ancillary function that takes an object of class smdi and produces a formatted and publication-ready gt table [39] which can be seamlessly exported to different file formats (e.g., .docx, .pdf, etc.) for reports or manuscripts.”*

- 7. In the article, the authors mention the use of synthetic datasets for their research. However, this approach may have some limitations, particularly in terms of generalizability and real-world data applicability. Synthetic datasets cannot fully represent the diversity and complexity of real-world data. It is recommended that consider validating their experimental results and comparing them with real-world data to ensure the accuracy and reliability of their findings.

[RESPONSE] We agree with the reviewer’s comment on the limited generalizability of synthetic datasets. We would like to emphasize that the synthetic example dataset that comes with the smdi package is provided for illustrative purposes and as a convenience for users to explore the functions and outputs of the package. This package allows users to implement, and streamline established statistical tests and diagnostic tools for missing data within a single package and user-friendly functions. We agree with the reviewer that the results and performance of the tests will vary by specific healthcare settings and the underlying complexity of the data. To get more experience with the package in real-world scenarios, additional work is under way to further evaluate the performance of these tools in different settings.