smdi: An R package to perform structural missing data investigations on partially observed confounders in real-world databases

**Authors**: Janick Weberpals1, Sudha R. Raman2, Pamela A. Shaw3, Hana Lee4, Bradley G. Hammill2, Sengwee Toh5, John G. Connolly5, Kimberly J. Dandreo5, Fang Tian4, Wei Liu4, Jie Li4, José J. Hernández-Muñoz4, Robert J. Glynn1, Rishi J. Desai1

Author affiliations:

1Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

2Department of Population Health Sciences, Duke University School of Medicine, Durham, NC

3Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA

4Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD

5Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA

**Correspondence:**

Janick Weberpals, RPh, PhD

Division of Pharmacoepidemiology and Pharmacoeconomics,

Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School,

1620 Tremont Street, Suite 3030-R, Boston, MA 02120, USA

Phone: 617-278-0932

Fax: 617-232-8602

Email: [jweberpals@bwh.harvard.edu](mailto:jweberpals@bwh.harvard.edu)

**Article type:**Application Note

**Word count:** xxxx words/2000 words

**Tables:** **x/2**

**Figures:** **x/3**

**Supplementary material:**

**Short running title**: An R package to perform structural missing data investigations

**Keywords:** Missing data, Confounder, EHR, R, Software, Real-World Evidence

**Funding:** This project was supported by Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).

**Disclosures/COI:** The FDA approved the study protocol, statistical analysis plan and reviewed and approved this manuscript. Coauthors from the FDA participated in the results interpretation and in the preparation and decision to submit the manuscript for publication. The FDA had no role in data collection, management, or analysis. The views expressed are those of the authors and not necessarily those of the US FDA. Janick Weberpals reports prior employment by Hoffmann-La Roche and previously held shares in Hoffmann-La Roche. Pamela Shaw is a named inventor on a patent licensed to Novartis by the University of Pennsylvania for an unrelated project. Sengwee Toh serves as a consultant for Pfizer, Inc. and TriNetX, LLC.. Robert J Glynn has received research funding through his employer from Amarin, Kowa, Novartis, and Pfizer. Dr. Desai reports serving as Principal Investigator on investigator-initiated grants to the Brigham and Women’s Hospital from Novartis, Vertex, and Bristol-Myers-Squibb on unrelated projects. All remaining authors report no disclosures or conflicts of interest.

**Analytical code and data sharing statement:** This manuscript was written using Quarto version 1.3.433 (<https://quarto.org/>) and R version 4.1.2. All R code, materials and depedency information can be found at <https://gitlab-scm.partners.org/drugepi/smdi-manuscript>. The R package presented in this study can be downloaded from CRAN via install.packages("smdi") or from <https://janickweberpals.gitlab-pages.partners.org/smdi/>.

**Acknowledgments:** We would like to thank all beta testers and attendees of the Division of Pharmacoepidemiology and Pharmacoeconomics Methods Incubator who gave valuable feedback on early versions of the smdi R package.

# Abstract

xxx words/150 words

|  |
| --- |
| Lay Summary |
| * … |

# Background and Significance

Electronic health records (EHR) are increasingly used to conduct real-world evidence (RWE) studies to complement evidence coming from randomized controlled trials (RCTs) [1,2]. Due to their detailed capture of clinical parameters such as vital signs, lab measurements, physician assessments and lifestyle factors, EHRs can significantly improve the ability to control for confounding and imbalances in prognostic factors between treatment groups, especially when linked to administrative claims databases [3]. However, such prognostic factors are often just partially observed which challenges the statistical analysis of the data and can result in severe bias when predicting or estimating treatment effects if not handled appropriately [4–6].

In order to make an informed decision about the most appropriate analytic approach, it is crucial to investigate and understand the potential patterns and mechanisms that underlie the partially observed confounder (POC) data (see definitions box) [7–9]. Usually these are not known for a given RWE study but general guidance papers and frameworks have suggested various routine diagnostics to investigate missing data patterns and mechanisms. These methods comprise standard procedures such as comparing baseline characteristics and outcomes between patients with and without the POC [10–14], checking the ability to predict missingness based on observed data [11] and assessing if causal relationships between variables and their missingness are recoverable based on available data [15] guided by directed acyclic graphs [16,17] or M-graphs [18]. However, these methods have so far been only described and tested in isolation from each other and no principled approach exists. In addition, the practical implementation of such diagnostics is time-consuming, tedious and consequently not often performed [19–21].

|  |
| --- |
| Definitions: Basic missing data taxonomies. |
| Patterns (adapted from Van Buuren [7])  * Monotone pattern: If Yj is the jth column in a dataset *Y*, a missing data pattern is said to be *monotone* if the variables Yj can be ordered such that if Yj is missing then all variables Yk with *k > j* are also missing. This can occur, for example, in longitudinal studies with drop-out or with clinical labs which are typically measured together as part of a lab panel (e.g., renal or liver panel). * Non-monotone pattern: If the pattern is not monotone, it is called *non-monotone* or *general*.  Mechanisms [11]  * **MCAR**: The missingness does not depend on any other observed or unobserved covariate(s). * **MCAR**: The missingness depends and can be explained by other observed covariates. * **MNAR**: The missingness depends on unoberserved covariate(s). For example, the missingness may be explained by other covariate(s) which is/are not observed in the underlying dataset (MNARunmeasured). The missingness can also just dependent on the actual value of the partially observed covariate itself (MNARvalue). |

In light of these limitations, we have recently developed and evaluated a principled approach combining a range of missing data diagnostics in an US EHR-claims database linkage [22]. The results of this large-scale study revealed that the combination of these diagnostics characterized missing data mechanisms well and provided helpful guidance for the appropriate choice of analytic methods to handle POC data (e.g., missing data imputation).

# Objective

To streamline and ease the implementation of these routine missing data diagnostics for confounder data in RWE studies, we developed the smdi (structural missing data investigations) R package [23]. The goal of smdi was to provide a toolbox of different diagnostics to analyze the potentially underlying missingness patterns and mechanisms of POCs in RWE studies.

# Materials and Methods

The smdi R package was written in R language (version 4.2.1). The package is available on CRAN (<https://cran.r-project.org/web/packages/smdi>){.uri}) and GitLab (<https://gitlab-scm.partners.org/janickweberpals/smdi>) and can be installed via the install.packages("smdi") command. To ensure the quality and robustness of the package and corresponding functions, we implemented comprehensive unit tests with a coverage of 95.81% and established automated R CMD checks [24] via continuous integration and deployment which are scheduled to execute on a daily basis. Additional resources such as documentation and vignettes are provided on the package website under <https://janickweberpals.gitlab-pages.partners.org/smdi>.

# Results

## Main Package Functions

[Figure 1](#fig-workflow) illustrates the recommended workflow and functions to systematically approach diagnostics on POCs.

The workflow is generally categorized into descriptives/pattern diagnostics and inferential diagnostics on potentially underlying missingness mechanisms. In this section, we cover the principles behind the main package functions, a selection of parameters a user can specify, the returned results and the results interpretation. Examples are illustrated using a simulated EHR dataset that is part of the package (more details under <https://janickweberpals.gitlab-pages.partners.org/smdi/articles/a_data_generation.html>). For all functions in the smdi package, a *dataframe* is expected (data parameter) as input with a format where one row represents one unique patient and the columns stand for the variables relevant for the study, i.e., the exposure, outcome, fully observed covariates and the POCs. Any non-informative columns, such as patient identifiers, date columns or zip codes should be dropped from the dataframe before calling the functions. Throughout all functions, the user has the option to specify a vector with the column name(s) of the POC(s) that should be investigated (covar parameter). If nothing is specified, all functions automatically consider any variable in the dataframe that exhibits at least one missing value.

### Descriptives and Pattern Diagnostics

As a first step to explore the missingness in new datasets, the smdi package provides a few basic functions to describe and summarize missingness across all covariates. The smdi\_summarize() function takes the dataframe as input and returns the amount and proportion of missing observations, which can also be stratified by a grouping variable (e.g., by an exposure or outcome variable). The smdi\_vis() function returns a corresponding bar chart plot (example [Figure 2](#fig-examples) a).

To visually inspect potential missing data patterns, we re-exported the gg\_miss\_upset() function of the naniar package [25]. This function uses a set visualization technique to visually infer potential (non-)monotone patterns based on the number of intersecting missing observations across all POCs [26]. For example, a monotone pattern would be visually evident if, for a set of two or more variables (e.g., lab1, lab2), all or the majority of missing records would be observed for both labs simultaneously (example [Figure 2](#fig-examples) b).

### Inferential Three Group Diagnostics

The core functions to infer potentially underlying missingness mechanisms are categorized into three group diagnostics based on their general analytic properties (**Table 1**).

#### Group 1 Diagnostics

The aim of the smdi\_asmd(), smdi\_hotelling() and smdi\_little() functions is to explore dissimilarities in patient characteristics between those with and without observed values for the POC. According to Rubin’s framework [8], when missingness is at random (MAR), it can be explained by observed covariates. Consequently, significant differences in patient characteristics would be expected under a MAR mechanism between strata of patients with and without the POC. If the missingness depends only on unobserved factors (missing not at random [MNAR]) or does not depend on either observed or unobserved covariates (missing completely at random [MCAR]), differences should not be observable.

To quantify and test such differences, the smdi\_asmd() function computes absolute standardized mean differences (ASMD) of observed patient characteristics [27–29]. The function returns an *asmd* object which displays an aggregated summary of the average or median ASMDs along with a corresponding range of minimum and maximum AMSDs observed. The object also returns a detailed ‘Table 1’ for each POC displaying the distributions of observed patient characteristics and resulting ASMDs between patients with and without an observed value for the POC. For a graphical visualization of this, the function returns a figure illustrating the ASMD for each compared patient characteristic (example [Figure 2](#fig-examples) c) [30].

The smdi\_hotelling() and smdi\_little() functions complement the smdi\_asmd() function by examining the differences in patient characteristics as a formal statistical hypothesis test. Hotelling’s test [12,31] formalizes this as a multivariate t-test for each POC, which means that smdi\_hotelling() returns a test statistic and a corresponding p-value for each each POC separately. In contrast, smdi\_little() [13,32] computes a single global chi-square test statistic and p-value across all POCs with the null hypothesis that the data is MCAR.

#### Group 2 Diagnostics

The group 2 diagnostics assesses the ability to predict missingness based on observed covariates via the smdi\_rf() function. This function trains and fits a random forest classification model [11,33] to predict the missing indicator of the POC given observed covariates as the predictors. If the resulting area under the receiver operating characteristic curve (AUC) is meaningfully higher than 0.5, this would give some evidence for MAR being the underlying missingness mechanism. In case of values close 0.5, this would rather indicate a random prediction and translate to a potential MCAR or MNAR mechanism.

The function returns an object of class *rf* which generically prints an overview of the AUC value of each POC. The AUC is based on the prediction made in the respective test dataset which is sampled as part of the function and for which the train-test split ratio can be chosen by the user (train\_test\_ratio parameter). The *rf* object further returns a graph for each POC displaying the relative importance of the predictors in the training dataset expressed as the mean decrease in accuracy (example [Figure 2](#fig-examples) d). This metric can be valuable for interpreting and identifying strong predictors of missingness. It quantifies how much the accuracy of the prediction (i.e., the ratio of correct predictions to the total number of predictions made) would decrease if we excluded a specific predictor from the prediction model. In case of inflated AUC values (>0.9), the function prompts a message to the user detailing the most important predictor. If another POC is identified as a perfect predictor, the presence of a monotone missing data pattern may be likely in which case it is recommended to exclude this POC and run the diagnostics for each POC independently.

#### Group 3 Diagnostics

The third group diagnostics with the smdi\_outcome() function examines the association of the missingness indicator of the POC and the outcome under study. The function will compute both a univariate model and a model adjusted for all other covariates included in the dataset. In previous simulations, we observed characteristic patterns for univariate and adjusted associations of the missing indicator and the outcome that matched simulated missingness mechanisms [22]. As one would expect, under a MCAR mechanism there was no difference in the outcome between patients with and without a value for the POC. Under MAR, given that missingness can be explained by observed covariates, a potential spurious association in the univariate model disappeared after adjustment. If the missingness followed any MNAR mechanism, an association was observed in both univariate and adjusted models.

Currently, smdi\_outcome() supports three outcome regression types: linear regression (*lm* [34]) for continuous outcomes, logistic regression (*glm* [34]) for binary outcomes and a Cox proportional hazards model (*coxph* [35]) for time-to-event outcomes. Besides the specification of the regression type, the user needs to specify the name of the column that contains the outcome using the *form\_lhs* parameter (e.g., Surv(eventtime, status) in case of a Cox model) and if resulting beta coefficients should be exponentiated or not. The function returns a table with univariate and adjusted beta coefficients and 95% confidence intervals for each POC.

#### smdi\_diagnose() to compute all three group diagnostics

Finally, the smdi\_diagnose() function enables a user to compute all three group diagnostics with just one function call.

# minimal example of a smdi\_diagnose() function call  
smdi\_diagnose(  
 data = smdi\_data,  
 covar = NULL, # NULL includes all covariates with at least one NA  
 model = "cox",  
 form\_lhs = "Surv(eventtime, status)",  
 n\_cores = 3 # number of cores to parallelize computations on  
 ) %>%   
 smdi\_style\_gt()

The function returns an object of class *smdi* containing a table with the results of all diagnostics for each specified POC and the global Little’s test p-value across all covariates (**Table 2**). 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 [36] which can be seamlessly exported to different file formats (e.g., .docx, .pdf, etc.) for reports or manuscripts.

# Discussion

Missing data are ubiquitous in real-world databases and may lead to bias if not handled appropriately. We developed smdi to streamline routine diagnostics of missing data. By cross-checking the resulting diagnostic parameters to expected estimates, based on the outlined theoretical assumptions and previous simulation results [22], the diagnostics can provide valuable insights into underlying missingness structures and help elucidate if analytic approaches such as missing imputation analyses are viable options (example shown in [Figure 3](#fig-results)).

The package also comes with limitations, such that the true underlying missingness generating mechanism can never be inferred with absolute certainty from the observed data. Hence, it’s important that diagnostic results are complemented with substantive expert knowledge to factor in how covariates are measured in routine care processes and contextualize potential reasons for missingness.

## Conclusions

The smdi package is a powerful and convenient tool to implement and carry out principled missing data diagnostics in RWE studies and improve the robustness of studies involving POCs by helping elucidate if certain analytic approaches are viable options for a given dataset.

# References

[1] United States Food and Drug Administration, Framework for FDA’s real world evidence program. Dec 2018. Accessed 6/30/2023., (n.d.). <https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf>.

[2] R.J. Desai, M.E. Matheny, K. Johnson, K. Marsolo, L.H. Curtis, J.C. Nelson, P.J. Heagerty, J. Maro, J. Brown, S. Toh, M. Nguyen, R. Ball, G. Dal Pan, S.V. Wang, J.J. Gagne, S. Schneeweiss, Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework, Npj Digital Medicine. 4 (2021). https://doi.org/[10.1038/s41746-021-00542-0](https://doi.org/10.1038/s41746-021-00542-0).

[3] A. Asfaw, M. Ascha, P. Yerram, S. Reiss, S. Brake, N. Wadé, SA27 Comparison of Comorbidity Indices Between Electronic Health Records (EHR) Derived Database and Claims Data Among Patients With Metastatic Breast Cancer, Value in Health. 25 (2022) S488. https://doi.org/[10.1016/j.jval.2022.09.2421](https://doi.org/10.1016/j.jval.2022.09.2421).

[4] M.H. Gorelick, Bias arising from missing data in predictive models, Journal of Clinical Epidemiology. 59 (2006) 1115–1123. https://doi.org/[10.1016/j.jclinepi.2004.11.029](https://doi.org/10.1016/j.jclinepi.2004.11.029).

[5] O.F. Ayilara, L. Zhang, T.T. Sajobi, R. Sawatzky, E. Bohm, L.M. Lix, Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry, Health and Quality of Life Outcomes. 17 (2019). https://doi.org/[10.1186/s12955-019-1181-2](https://doi.org/10.1186/s12955-019-1181-2).

[6] R.H.H. Groenwold, O.M. Dekkers, Missing data: The impact of what is not there, European Journal of Endocrinology. 183 (2020) E7–E9. https://doi.org/[10.1530/eje-20-0732](https://doi.org/10.1530/eje-20-0732).

[7] S. Van Buuren, Flexible imputation of missing data, CRC press, 2018. <https://stefvanbuuren.name/fimd/missing-data-pattern.html>.

[8] D.B. RUBIN, Inference and missing data, Biometrika. 63 (1976) 581–592. https://doi.org/[10.1093/biomet/63.3.581](https://doi.org/10.1093/biomet/63.3.581).

[9] R.J. Little, D.B. Rubin, Statistical analysis with missing data, John Wiley & Sons, 2019.

[10] K.J. Lee, K.M. Tilling, R.P. Cornish, R.J.A. Little, M.L. Bell, E. Goetghebeur, J.W. Hogan, J.R. Carpenter, Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework, Journal of Clinical Epidemiology. 134 (2021) 79–88. https://doi.org/[10.1016/j.jclinepi.2021.01.008](https://doi.org/10.1016/j.jclinepi.2021.01.008).

[11] A. Sondhi, J. Weberpals, P. Yerram, C. Jiang, M. Taylor, M. Samant, S. Cherng, 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. (2023). https://doi.org/[10.1002/psp4.12998](https://doi.org/10.1002/psp4.12998).

[12] H. Hotelling, The Generalization of Student’s Ratio, The Annals of Mathematical Statistics. 2 (1931) 360–378. https://doi.org/[10.1214/aoms/1177732979](https://doi.org/10.1214/aoms/1177732979).

[13] R.J.A. Little, A Test of Missing Completely at Random for Multivariate Data with Missing Values, Journal of the American Statistical Association. 83 (1988) 1198–1202. https://doi.org/[10.1080/01621459.1988.10478722](https://doi.org/10.1080/01621459.1988.10478722).

[14] A. Pedersen, E. Mikkelsen, D. Cronin-Fenton, N. Kristensen, T.M. Pham, L. Pedersen, I. Petersen, Missing data and multiple imputation in clinical epidemiological research, Clinical Epidemiology. Volume 9 (2017) 157–166. https://doi.org/[10.2147/clep.s129785](https://doi.org/10.2147/clep.s129785).

[15] P. Madley-Dowd, R. Hughes, K. Tilling, J. Heron, The proportion of missing data should not be used to guide decisions on multiple imputation, Journal of Clinical Epidemiology. 110 (2019) 63–73. https://doi.org/[10.1016/j.jclinepi.2019.02.016](https://doi.org/10.1016/j.jclinepi.2019.02.016).

[16] K.J. Lee, J.B. Carlin, J.A. Simpson, M. Moreno-Betancur, Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification, International Journal of Epidemiology. (2023). https://doi.org/[10.1093/ije/dyad008](https://doi.org/10.1093/ije/dyad008).

[17] M. Moreno-Betancur, K.J. Lee, F.P. Leacy, I.R. White, J.A. Simpson, J.B. Carlin, Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies, American Journal of Epidemiology. 187 (2018) 2705–2715. https://doi.org/[10.1093/aje/kwy173](https://doi.org/10.1093/aje/kwy173).

[18] K. Mohan, J. Pearl, Graphical Models for Processing Missing Data, Journal of the American Statistical Association. 116 (2021) 1023–1037. https://doi.org/[10.1080/01621459.2021.1874961](https://doi.org/10.1080/01621459.2021.1874961).

[19] O.U. Carroll, T.P. Morris, R.H. Keogh, How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review, BMC Medical Research Methodology. 20 (2020). https://doi.org/[10.1186/s12874-020-01018-7](https://doi.org/10.1186/s12874-020-01018-7).

[20] A.M. Wood, I.R. White, S.G. Thompson, Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals, Clinical Trials. 1 (2004) 368–376. https://doi.org/[10.1191/1740774504cn032oa](https://doi.org/10.1191/1740774504cn032oa).

[21] O. Harel, J. Pellowski, S. Kalichman, Are We Missing the Importance of Missing Values in HIV Prevention Randomized Clinical Trials? Review and Recommendations, AIDS and Behavior. 16 (2012) 1382–1393. https://doi.org/[10.1007/s10461-011-0125-6](https://doi.org/10.1007/s10461-011-0125-6).

[22] J. Weberpals, S.R. Raman, P.A. Shaw, H. Lee, B.G. Hammill, S. Toh, J.G. Connolly, K.J. Dandreo, F. Tian, W. Liu, J. Li, J.J. Hernández-Muñoz, R.J. Glynn, R.J. Desai, A principled approach to characterize and analyze partially observed confounder data from electronic health records, Submitted. (2023).

[23] J. Weberpals, Smdi: Perform structural missing data investigations, (2023). <https://CRAN.R-project.org/package=smdi>.

[24] H. Wickham, J. Bryan, R packages, " O’Reilly Media, Inc.", 2023.

[25] N. Tierney, D. Cook, Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations, 105 (2023). https://doi.org/[10.18637/jss.v105.i07](https://doi.org/10.18637/jss.v105.i07).

[26] R.A. Ruddle, M. Adnan, M. Hall, Using set visualisation to find and explain patterns of missing values: a case study with NHS hospital episode statistics data, BMJ Open. 12 (2022) e064887. https://doi.org/[10.1136/bmjopen-2022-064887](https://doi.org/10.1136/bmjopen-2022-064887).

[27] P. Schober, T.R. Vetter, Correct Baseline Comparisons in a Randomized Trial, Anesthesia & Analgesia. 129 (2019) 639. https://doi.org/[10.1213/ane.0000000000004211](https://doi.org/10.1213/ane.0000000000004211).

[28] P.C. Austin, An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies, Multivariate Behavioral Research. 46 (2011) 399–424. https://doi.org/[10.1080/00273171.2011.568786](https://doi.org/10.1080/00273171.2011.568786).

[29] K. Yoshida, A. Bartel, Tableone: Create ’table 1’ to describe baseline characteristics with or without propensity score weights, (2022). <https://CRAN.R-project.org/package=tableone>.

[30] H. Wickham, ggplot2: Elegant graphics for data analysis, (2016). <https://ggplot2.tidyverse.org>.

[31] J. Curran, T. Hersh, Hotelling: Hotelling’s t^2 test and variants, (2021). <https://CRAN.R-project.org/package=Hotelling>.

[32] N. Tierney, D. Cook, Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations, 105 (2023). https://doi.org/[10.18637/jss.v105.i07](https://doi.org/10.18637/jss.v105.i07).

[33] A. Liaw, M. Wiener, Classification and regression by randomForest, 2 (2002) 18–22. <https://CRAN.R-project.org/doc/Rnews/>.

[34] R Core Team, R: A language and environment for statistical computing, (2022). <https://www.R-project.org/>.

[35] T.M. Therneau, A package for survival analysis in r, (2023). <https://CRAN.R-project.org/package=survival>.

[36] R. Iannone, J. Cheng, B. Schloerke, E. Hughes, A. Lauer, J. Seo, Gt: Easily create presentation-ready display tables, (2023). <https://CRAN.R-project.org/package=gt>.

# Tables

Table 1: Overview of the main functions in smdi to characterize potential underlying missingness mechanisms.

| Function | Description | Generic S3 print() output | Object output | Interpretation |
| --- | --- | --- | --- | --- |
| Group 1 Diagnostics - Comparing the distribution of observed covariates between patients with versus without a value for the partially observed covariate | | | | |
| **smdi\_asmd()** | Computes the absolute standardized mean differences (ASMD) of patient characteristics between patients with versus without a value for the partially observed covariate(s) | Aggregated summary table of the average/median and minimum/maximum ASMD range for all specified partially observed covariates | - Detailed *Table 1* illustrating distributions and individual ASMD for each compared patient characteristic  - ggplot2 graph illustrating the individual ASMD for each compared patient characteristic in descending order  - Aggregate summary of the average/median and minimum/maximum ASMD range for the selected partially observed covariate | - ASMD < 0.1: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR)  - ASMD > 0.1: imbalances in observed patient characteristics; missingness may be likely at random (~MAR) |
| **smdi\_hotelling()** | Computes Hotelling's multivariate t-test for each partially observed covariate, examining patient differences conditional on having an observed covariate value or not. | Aggregated summary table of the Hotelling's test p-values for all specified partially observed covariates | Detailed Hotelling test statistics | High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR) |
| **smdi\_little()** | Computes a single global chi-square test statistic across all partially observed covariates with a null hypothesis that the data is missing completely at random. | Detailed Little's test statistics | Detailed Little's test statistics | High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR) |
| Group 2 Diagnostics - Assessing the ability to predict missingness based on observed covariates | | | | |
| **smdi\_rf()** | Trains and fits a random forest classification model to assess the ability to predict missingness indicator for the partially observed covariate(s). | Aggregated summary table with the area under the receiver operating characteristic curve (AUC) value for each partially observed covariate | - Individual AUC value  - ggplot2 figure illustrating the variable importance for the prediction made expressed by the mean decrease in accuracy per predictor | - AUC values ~ 0.5 indicate completely random or not at random prediction (~MCAR, ~MNAR)  - Values meaningfully above 0.5 indicate stronger relationships between covariates and missingness (~MAR) |
| Group 3 Diagnostics - Evaluates whether missingness of a covariate is associated with the outcome | | | | |
| **smdi\_outcome()** | Fits outcome model (linear, logistic or proportional hazards depending on the outcome under study) with the missingness indicator of the partially observed covariate(s). The estimates are computed both as a univariate model (just considering the missingness indicator) and an adjusted model with all covariates in the dataset. | Aggregated summary table with the univariate and adjusted estimate for each partially observed covariate | Aggregated summary table with the univariate and adjusted estimate for each partially observed covariate | - No association in either univariate or adjusted model and no meaningful difference in the log HR after full adjustment (~MCAR).  - Association in univariate but not fully adjusted model (~MAR).  - Meaningful difference in the log HR also after full adjustment (~MNAR). |

Table 2: Example output of the smdi\_diagnose() function applied to the examplary smdi\_data dataset.

| Covariate | ASMD (min/max)*1* | p Hotelling*1* | AUC*2* | beta univariate (95% CI)*3* | beta (95% CI)*3* |
| --- | --- | --- | --- | --- | --- |
| ecog\_cat | 0.029 (0.003, 0.071) | 0.783 | 0.510 | -0.06 (95% CI -0.16, 0.03) | -0.06 (95% CI -0.16, 0.03) |
| egfr\_cat | 0.243 (0.010, 0.485) | <.001 | 0.629 | 0.06 (95% CI -0.03, 0.15) | -0.01 (95% CI -0.10, 0.09) |
| pdl1\_num | 0.062 (0.019, 0.338) | <.001 | 0.516 | 0.12 (95% CI 0.01, 0.23) | 0.11 (95% CI -0.00, 0.22) |
| p little: <.001, Abbreviations: ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, beta = beta coefficient, CI = Confidence interval, max = Maximum, min = Minimum | | | | | |
| *1*Group 1 diagnostic: Differences in patient characteristics between patients with and without covariate | | | | | |
| *2*Group 2 diagnostic: Ability to predict missingness | | | | | |
| *3*Group 3 diagnostic: Assessment if missingness is associated with the outcome (univariate, adjusted) | | | | | |

# Figures

|  |
| --- |
| Figure 1: Overview of all smdi functions and suggested workflow to conduct missing data diagnostics. \*gg\_miss\_upset() and md.pattern() are re-exports of the naniar and mice package, respectively. |

|  |
| --- |
| Figure 2: Exemplary visual outputs of the a) smdi\_vis(), b) gg\_miss\_upset(), c) smdi\_asmd() and d) smdi\_rf() functions, respectively. |

|  |
| --- |
| Figure 3: Example of how smdi diagnostics can be applied to compute and compare diagnostic parameters of partially observed covariates to expected parameters of common missingness mechanisms based on a former large-scale simulation study [22]. |