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

**Authors**: Janick Weberpals1, RPh, PhD, Sudha R. Raman2, PhD, Pamela A. Shaw3, PhD, MS, Hana Lee4, PhD, Bradley G. Hammill2, DrPH, Sengwee Toh5, ScD, John G. Connolly5, ScD, Kimberly J. Dandreo5, MS, Fang Tian6, PhD, Wei Liu6, PhD, Jie Li6, PhD, José J. Hernández-Muñoz6, PhD, Robert J. Glynn1, PhD, ScD, Rishi J. Desai1, PhD

Author affiliations:

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

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

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

4 Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD

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

6 Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD

**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:** 2,514 words

**Tables:** 2

**Figures:** 3

**Supplementary material:** Supplementary Methods

**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 depedencies can be found at <https://gitlab-scm.partners.org/drugepi/smdi-manuscript> or <https://github.com/janickweberpals/smdi-manuscript>. The R package presented in this study can be downloaded from the comprehensive R archive network (CRAN) via install.packages("smdi") (version 0.2.2 at time of manuscript submission) or from [https://janickweberpals.gitlab-pages.partners.org/smdi](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

150 words/150 words

**Objectives**

Partially observed confounder data pose a major challenge in statistical analyses aimed to inform causal inference using electronic health records (EHR). While analytic approaches such as imputation are available, assumptions on underlying missingness patterns and mechanisms must be verified. We aimed to develop a toolkit to streamline missing data diagnostics to guide choice of analytic approaches based on meeting necessary assumptions.

**Materials and Methods**

We developed the smdi (structural missing data investigations) R package based on results of a previous simulation study which considered structural assumptions of common missing data mechanisms in EHR.

**Results**

smdi enables users to run principled missing data investigations on partially observed confounders and implement functions to visualize, describe, and infer potential missingness patterns and mechanisms based on observed data.

**Conclusions**

The smdi R package is freely available on CRAN and can provide valuable insights into underlying missingness patterns and mechanisms and thereby help improve the robustness of real-world evidence studies.

|  |
| --- |
| Lay Summary |
| Contrary to clinical trials, which are designed to collect data for research purposes in a harmonized manner, real-world data are typically generated for administrative (e.g., health insurance claims for billing purposes) or clinical documentation (e.g., electronic health records) purposes. Hence, confounders and prognostic factors, which need to be balanced between treatment groups to infer causal treatment effects, are often unavailable for all patients at all necessary time points, resulting in missing data. It is important to investigate potential missingness patterns and mechanisms to determine if assumptions for common analytic methods such as complete case analysis or imputations hold to avoid bias in resulting treatment effect estimates. Here, we present the smdi R package, which enables researchers to conduct such principled missing data investigations on partially observed confounders. The smdi package implements functions to visualize and describe missing data and to infer potential missingness patterns and mechanisms based on observed data. |

# Background and Significance

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 initiative [1], linking claims databases and electronic health records (EHR) to generate real-world evidence (RWE) and complement data from randomized controlled trials (RCTs) [1,2]. Due to their capture of clinical details, EHR can significantly improve the ability to mitigate imbalances in prognostic factors between treatment groups, especially when linked to administrative claims databases [3]. At the moment, substantial efforts focusing on the linkage of claims databases and EHR are underway, for instance, in the FDA Sentinel Initiative [1]. However, prognostic factors coming from EHR are often only partially observed, posing a challenge to the statistical analysis and potentially leading to bias in treatment effect estimates if not handled appropriately [4–6].

In order to inform decisions about the most appropriate analytic approach, it is useful to investigate the potential patterns and mechanisms that underlie the partially observed confounder (POC) data (see definitions box) [7–9]. Existing guidance 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 [11] and assessing if causal relationships between variables and their missingness are recoverable based on available data [15] using directed acyclic graphs [16,17] or M-graphs [18]. However, these methods have so far only been described and tested in isolation from each other and no unified principled approach exists. In addition, the practical implementation of such diagnostics is time-consuming and consequently infrequently 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. * Non-monotone pattern: If the pattern is not monotone, it is called *non-monotone* or *general*.  Mechanisms [11]  * **Missing completely at random (MCAR)**: The missingness does not depend on any other observed or unobserved covariate(s). * **Missing at random (MAR)**: The missingness depends and can be explained by other observed covariates. * **Missing not at random (MNAR)**: The missingness depends on unobserved 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 depend on the actual value of the partially observed covariate itself (MNARvalue). |

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]. The results of this large-scale study revealed that the combination of these diagnostics effectively identified underlying mechanisms and provided helpful guidance for the choice of appropriate analytic methods to handle POC data.

# Objective

To streamline the implementation of routine missing data diagnostics for POC data in RWE studies, we developed the smdi (structural missing data investigations) R package [25].

# Materials and Methods

The smdi R package was written in R language (version 4.2.1). The package is available on the comprehensive R archive network (<https://cran.r-project.org/web/packages/smdi>) and GitLab (<https://gitlab-scm.partners.org/janickweberpals/smdi>) and can be installed via install.packages("smdi"). To ensure quality, we implemented comprehensive unit tests with a coverage of 95.81% and established automated R CMD checks [26] via continuous integration and deployment. Additional resources such as documentation and vignettes are provided on the package website: <https://janickweberpals.gitlab-pages.partners.org/smdi>.

# Results

## Main Package Functions

[Figure 1](#fig-workflow) illustrates the recommended workflow 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 users can specify, the returned results and how these can be interpreted. Examples are illustrated using a synthetic dataset that is part of the package and simulates an oncology cohort with a binary exposure, a time-to-event outcome and several baseline confounders and prognostic covariates including three POCs (EGFR and PD-L1 [biomarkers] and ECOG [performance score]) following a MAR, MNAR and MCAR mechanism, respectively (more details: <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 represent relevant variables, i.e., exposure, outcome, fully observed covariates and the POCs. Any non-informative columns, e.g., patient identifiers, should be dropped from the dataframe before calling the functions. Throughout all functions, users have 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.

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()

### 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 returns the amount and proportion of missing observations, which can also be stratified by a grouping 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 [27]. 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 [28]. For example, a monotone pattern could be visually evident if, for a set of two or more lab variables which are typically measured together as part of a lab panel (e.g. renal or liver panel), the missingness of one lab is indicative of the missingness in the other lab and hence all or the majority of combinations of cells are missing together (example [Figure 2](#fig-examples) b). 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.

### 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 such differences, the smdi\_asmd() function computes absolute standardized mean differences (ASMD) of observed patient characteristics [30–32]. The function returns an *asmd* object which displays an aggregated summary of the average or median ASMD along with a corresponding range of minimum and maximum ASMDs for each POC, respectively. The object also returns detailed ‘Table 1’s’ and plots [33] for each POC displaying the distributions of observed covariates and resulting ASMDs between patients with and without an observed value for the POC (example [Figure 2](#fig-examples) c).

The smdi\_hotelling() and smdi\_little() functions complement the smdi\_asmd() function by examining differences in patient characteristics as a formal statistical hypothesis test. Hotelling’s test [12,34] formalizes this as a multivariate t-test, which means that smdi\_hotelling() returns a test statistic and p-value for each POC. In contrast, smdi\_little() [13,27] computes a single global chi-square test statistic and p-value across all POCs with the null hypothesis that the data are (globally) MCAR.

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](#fig-results) 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](#fig-examples) c, [Figure 3](#fig-results)).

#### Group 2 Diagnostics

Group 2 diagnostics assess the ability to predict missingness based on observed covariates via the smdi\_rf() function. This function trains and fits a random forest classification model [11,35] to predict the missing indicator of each POC given exposure, outcome, follow-up time, and covariates plus missingness indicator for other POC 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/against MCAR being the underlying missingness mechanism. In case of values close 0.5, this would indicate the model is unable to discriminate missing vs. observed values based on available data; this could be due to a mechanism that is close to MCAR or one where the missingness is associated with unmeasured data (MNAR).

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, number of trees and CPU cores to parallelize over can be specified (train\_test\_ratio, ntree and n\_cores parameter, respectively) [35,36]. 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 model. In case of inflated AUC values (>0.9), the function prompts a message to the user reporting the most important predictor. If in such a scenario missingness in 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 run the diagnostics for each POC independently rather than jointly.

[Figure 3](#fig-results) (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.

#### Group 3 Diagnostics

The third group of diagnostics with the smdi\_outcome() function examines the association of the missingness indicator of the POC and the outcome under study. The function computes both a univariate model and a model adjusted for all other covariates in the dataset. In simulations, we discerned distinct patterns in both univariate and adjusted associations between the missing indicator and the outcome, closely mirroring simulated missingness mechanisms ([Figure 3](#fig-results) top) [22]. As expected, under a MCAR mechanism the simulation suggested no difference in the outcome between patients with and without a value for the POC. Under MAR, given that missingness can be sufficiently explained by observed covariates, a spurious association in the univariate model disappeared after adjustment. If the missingness followed any MNAR mechanism, an association was observed regardless of adjustment.

smdi\_outcome() supports multiple outcome regression types including linear regression (*lm* [37]) for continuous outcomes, Cox proportional hazards model (*coxph* [38]) 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). The function returns a table with univariate and adjusted beta coefficients and 95% confidence intervals for each POC.

Demonstrating the utilization of smdi\_outcome() using the example dataset, the derived logHR coefficients for the missingness indicators of the POCs ([Figure 3](#fig-results) 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.

#### 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](#fig-results), [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.

# Discussion

Missing data are ubiquitous in RWE studies involving EHR and may introduce bias if not handled appropriately. To address this issue, we developed the smdi R package to streamline routine diagnostics of missing data.

The package does come with certain limitations, such that the true underlying mechanism causing the missing data can never be inferred with absolute certainty from the observed data. Therefore, it’s important to complement diagnostic results with substantive expert knowledge to factor in how covariates are measured in routine care and contextualize potential reasons for missingness. This collaborative approach allows for a contextualized understanding of potential causes for missing data in EHR.

## Conclusions

The smdi R package is a powerful and convenient tool to implement principled routine missing data diagnostics in RWE studies. This will improve the robustness of studies involving POCs by elucidating if certain analytic approaches are viable for a given dataset.

# References

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 Digital Medicine*. 2021;4. doi: [10.1038/s41746-021-00542-0](https://doi.org/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.](https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf)

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

4 Gorelick MH. [Bias arising from missing data in predictive models](https://doi.org/10.1016/j.jclinepi.2004.11.029). *Journal of Clinical Epidemiology*. 2006;59:1115–23.

5 Ayilara OF, Zhang L, Sajobi TT, *et al.* 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*. 2019;17. doi: [10.1186/s12955-019-1181-2](https://doi.org/10.1186/s12955-019-1181-2)

6 Groenwold RHH, Dekkers OM. [Missing data: The impact of what is not there](https://doi.org/10.1530/eje-20-0732). *European Journal of Endocrinology*. 2020;183:E7–9.

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

8 RUBIN DB. [Inference and missing data](https://doi.org/10.1093/biomet/63.3.581). *Biometrika*. 1976;63:581–92.

9 Little RJ, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons 2019.

10 Lee KJ, Tilling KM, Cornish RP, *et al.* [Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework](https://doi.org/10.1016/j.jclinepi.2021.01.008). *Journal of Clinical Epidemiology*. 2021;134:79–88.

11 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. doi: [10.1002/psp4.12998](https://doi.org/10.1002/psp4.12998)

12 Hotelling H. [The Generalization of Student’s Ratio](https://doi.org/10.1214/aoms/1177732979). *The Annals of Mathematical Statistics*. 1931;2:360–78.

13 Little RJA. [A Test of Missing Completely at Random for Multivariate Data with Missing Values](https://doi.org/10.1080/01621459.1988.10478722). *Journal of the American Statistical Association*. 1988;83:1198–202.

14 Pedersen A, Mikkelsen E, Cronin-Fenton D, *et al.* [Missing data and multiple imputation in clinical epidemiological research](https://doi.org/10.2147/clep.s129785). *Clinical Epidemiology*. 2017;Volume 9:157–66.

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

16 Lee KJ, Carlin JB, Simpson JA, *et al.* Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. *International Journal of Epidemiology*. Published Online First: 13 February 2023. doi: [10.1093/ije/dyad008](https://doi.org/10.1093/ije/dyad008)

17 Moreno-Betancur M, Lee KJ, Leacy FP, *et al.* [Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies](https://doi.org/10.1093/aje/kwy173). *American Journal of Epidemiology*. 2018;187:2705–15.

18 Mohan K, Pearl J. [Graphical Models for Processing Missing Data](https://doi.org/10.1080/01621459.2021.1874961). *Journal of the American Statistical Association*. 2021;116:1023–37.

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

20 Wood AM, White IR, Thompson SG. [Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals](https://doi.org/10.1191/1740774504cn032oa). *Clinical Trials*. 2004;1:368–76.

21 Harel O, Pellowski J, Kalichman S. [Are We Missing the Importance of Missing Values in HIV Prevention Randomized Clinical Trials? Review and Recommendations](https://doi.org/10.1007/s10461-011-0125-6). *AIDS and Behavior*. 2012;16:1382–93.

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](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839563/). *AMIA Annual Symposium Proceedings*. 2006;2006:1044.

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

25 Weberpals J. [Smdi: Perform structural missing data investigations](https://CRAN.R-project.org/package=smdi). Published Online First: 2023.

26 Wickham H, Bryan J. *R packages*. " O’Reilly Media, Inc." 2023.

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

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

29 Buuren S van, Groothuis-Oudshoorn K. [Mice: Multivariate imputation by chained equations in r](https://doi.org/10.18637/jss.v045.i03). 2011;45:1–67.

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

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

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

33 Wickham H. [ggplot2: Elegant graphics for data analysis](https://ggplot2.tidyverse.org). Published Online First: 2016.

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

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

36 [Breiman L. *Machine Learning*. 2001;45:5–32.](https://doi.org/10.1023/a:1010933404324)

37 R Core Team. [R: A language and environment for statistical computing](https://www.R-project.org/). Published Online First: 2022.

38 Therneau TM. [A package for survival analysis in r](https://CRAN.R-project.org/package=survival). Published Online First: 2023.

39 Iannone R, Cheng J, Schloerke B, *et al.* [Gt: Easily create presentation-ready display tables](https://CRAN.R-project.org/package=gt). Published Online First: 2023.

# 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 are 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  - Estimated out-of-bag (OOB) error | - 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, glm 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. In this example, ECOG performance score (ecog\_cat) shows no imbalances in patient characteristics between patient with and without an observed value (absolute standardized mean difference [ASMD] 0.029, p[Hotelling] 0.783, group 1 diagnostic). Additionally missingness cannot be predicted well (AUC = 0. 510, group 2 diagnostic) and no difference in the outcome can be observed between patients with and without ecog\_cat (log HR -0.06 [95% CI -0.16; 0.03], group 3 diagnostic). Accordingly, the missingness diagnostics indicate that ECOG follows a missing completely at random missingness (MCAR) mechanism. Similarly, the EGFR (egfr\_cat) and PD-L1 (pdl1\_num) biomarker variables can be interpreted as following a missing at random (MAR) and missing not at random value (MNARvalue) mechanism. See also [Figure 3](#fig-results).

| 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. Sub-figure a) displays the proportion of missing observations for each partially observed covariate stratified by exposure. The upset plot in sub-figure b) demonstrates how a monotone missingness pattern between partially observed covariates can be visually inspected using a set visualization technique [28]. Sub-figure c) illustrates absolute standardized mean differences (ASMD) in patient characteristics between patients with and without a value observed for the PD-L1 (pdl1\_num) biomarker as a measure of imbalance. Sub-figure d) demonstrates the variable importance of fully observed covariates for predicting missingness in the partially oberserved ECOG performance score variable (ecog\_cat). |

|  |
| --- |
| 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]. |