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)

**Word count:** xxxx words

**Tables:**

**Figures:**

**Supplementary material:**

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

**Keywords:** Missing data, Confounder, EHR, R, Software,

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

**Disclosures/COI:res/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 and materials can be found at <https://gitlab-scm.partners.org/drugepi/smdi-manuscript>. Detailed information on packages and versions can be found in the renv.lock file in the linked repository. The R package and data presented in this study can be downloaded 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

|  |
| --- |
| Highlights |
| * … |

# Background

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 to arrive at unbiased estimates, it is crucial to investigate and understand the potential patterns and mechanisms that underlie the partially observed confounder data [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 distributions of baseline characteristics and outcomes between patients with and without the partially observed covariate(s) [10–13], 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 [14] and guided by directed acyclic graphs [15,16] or M-graphs [17]. 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 all of these diagnostics is time-consuming, tedious and is consequently not often performed in both RWE and RCTs [18–20].

To overcome these shortcomings, we [21] 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 study revealed that the combination of these diagnostics characterized simulated missing data mechanisms well and provided helpful insights for the appropriate choice of analytic methods to handle the partially observed confounder data (e.g., missing data imputation).

To streamline and ease the implementation of these routine missing data diagnostics for confounder data in RWE studies, we here present and demonstrate the use of the smdi (structural missing data investigations) R package [23]. To that end, we first provide some brief background on the theoretical assumptions that underlie the diagnostic functions of this package. We then give an overview of the implemented functions and demonstrate an end-to-end workflow application of the package with a hypothetical RWE study using simulated oncology EHR dataset that comes bundled with the package.

# Methods

## Problem formalization

As opposed to clinical trials, which are designed to collect data for research purposes in a harmonized manner, real-world data (RWD) are typically generated for administrative purposes (e.g., health insurance claims for billing purposes) or clinical documentation (e.g., EHR). Hence, confounders and prognostic factors, which need to be balanced between treatment groups in RWE studies are usually not available for all patients at all neccessary time points [24]. If the underlying mechanism for the missingness of such covariates is informative towards the outcome under study, e.g., patients who have a certain biomarker measured have a significantly increased risk if experiencing the outcome, this can lead to bias in the resulting effect estimates for the studied treatments under common missing data approaches like complete case analysis or imputation [6]. Hence, it is of utmost importance to investigate the potential patterns and mechanisms to know if assumptions for specific missing data approaches hold [25,26].

## Theoretical background

In our previous work [22] and for the purpose of the smdi R package, we categorized the implemented principled missing data investigations into three groups of diagnostics, which will be discussed in this section (Table 1).

Functions overview

….

Data generation for hypothetical study

….

# Results

# Discussion

# 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.

[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.D. 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. (accepted), CPT Pharmacometrics Syst Pharmacol. (2023).

[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] 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).

[15] 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).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] Sentinel Initiative Workstream, Approaches to handling partially observed confounder data from electronic health records (EHR) in non-randomized studies of medication outcomes. Accessed 7/14/2023., (n.d.). <https://www.sentinelinitiative.org/methods-data-tools/methods/approaches-handling-partially-observed-confounder-data-electronic-health>.

[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] S. Toh, L.A. García Rodríguez, M.A. Hernán, Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics, Pharmacoepidemiology and Drug Safety. 21 (2012) 13–20. https://doi.org/[10.1002/pds.3248](https://doi.org/10.1002/pds.3248).

[25] 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).

[26] K.J. Lee, J.A. Simpson, Introduction to multiple imputation for dealing with missing data, Respirology. 19 (2013) 162–167. https://doi.org/[10.1111/resp.12226](https://doi.org/10.1111/resp.12226).

# Tables

Table 1: Overview of the main functions in smdi.

| 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^2^ 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 predic 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). |

# Figures