Supplementary Material

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

Authors: Janick Weberpals[1], RPh, PhD, Sudha R. Raman[2], PhD, Pamela A. Shaw[3], PhD, MS, Hana Lee[4], PhD, Bradley G. Hammill[2], DrPH, Sengwee Toh[5], ScD, John G. Connolly[5], ScD, Kimberly J. Dandreo[5], MS, Fang Tian[6], PhD, Wei Liu[6], PhD, Jie Li[6], PhD, José J. Hernández-Muñoz[6], PhD, Robert J. Glynn[1], PhD, ScD, Rishi J. Desai[1], 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

1 Supplementary Methods

1.1 Missingness assumptions

For the design and validation of the `smdi` R package functions, we employed comprehensive simulations and a real-world example [1]. These simulations followed structural assumptions on realistic missingness generating mechanisms that could be expected in electronic health record (EHR) data. The below directed acyclic diagrams (DAG) [2,3], where a directed edge represents a causal effect of one variable on another variable, outline and illustrate these assumptions along with a brief explanations and exemplary real-world EHR scenarios. Since the `smdi` package was developed with focus on partially observed confounders (as opposed to missingness in exposure or outcome), we simplified the below DAGs by leaving out the nodes for exposure and outcome.

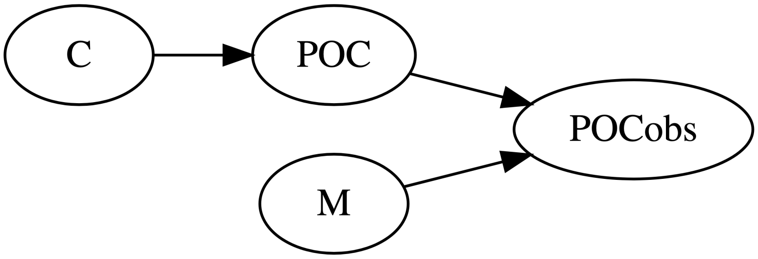
Notation: C = Fully observed confounder, POC = True values of the partially observed confounder, POCobs = Observed portion of the partially observed confounder, M = Missingness of POC with M=0 fully observed and M=1 fully missing, U = Unobserved confounder.

1.1.1 Missing completely at random (MCAR)

The missingness (M) of the partially observed confounder (POC) is independent of any observed and unobserved confounders. That is, no edge directs to M and the missingness is completely at random (MCAR).

Real-world example: A machine breaks that is used to measure and analyze a certain biomarker.

Supplementary Figure 1: Directed acyclic graph displaying a missing completely at random (MCAR) scenario.

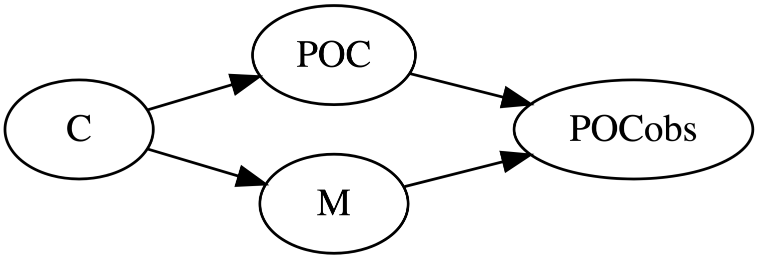


1.1.2 Missing at random (MAR)

The missingness (M) of the partially observed confounder (POC) can be explained by one or multiple fully observed confounders (C). That is, there is an edge between C to M and the missingness is at random (MAR).

Real-world example: Older patients are more likely to receive a certain biomarker test and age is a fully observed confounder which is measured for every patient in the dataset.

Supplementary Figure 2: Directed acyclic graph displaying a missing completely at random scenario.

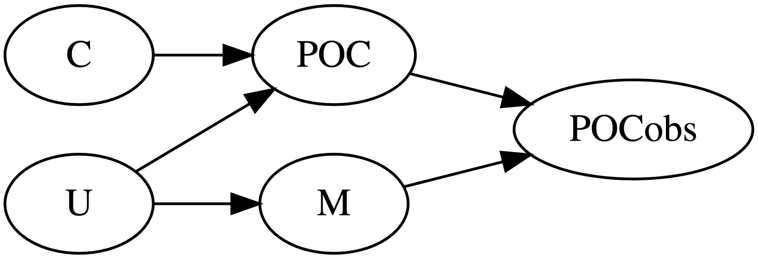


1.1.3 Missing not at random - unmeasured (MNAR$_{\text{unmeasured}}$)

The missingness (M) of the partially observed confounder (POC) can only be explained by one or multiple confounder(s) which are not observed in the dataset.

Real-world example: Patients with a certain performance status, which not observed in the dataset, are more likely to receive a lab test.

Supplementary Figure 3: Directed acyclic graph displaying a missing not at random (unmeasured) scenario.

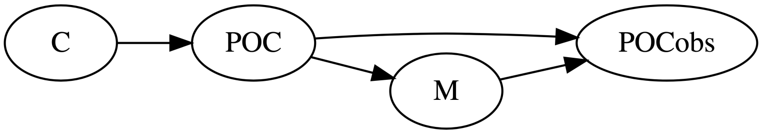


1.1.4 Missing not at random - value (MNAR$_{value}$)

The missingness (M) of the partially observed confounder (POC) depends on the true value of the partially observed confounder itself.

Real-world example: Patients with a history of normal biomarker values are systematically less likely to be tested again for the same biomarker.

Supplementary Figure 4: Directed acyclic graph displaying a missing not at random (value) scenario.



1.2 Missingness characterization

1.2.1 Group 1 diagnostics

The featured group 1 diagnostics in `smdi` focus on investigating potential differences in the distribution of characteristics between patients with and without an observed value for a partially observed confounder. To that end, different analytic approaches are leveraged which are described for each function below.

- `smdi_hotelling()`: This function computes a two-sample Hotelling's T-squared test for the difference in two multivariate means based on the Hotelling's T-Square test statistic [4,5]. It assesses whether there is a statistically significant difference between the means of two groups in multivariate data. This test is an extension of the univariate t-test that examines for two groups (i.e., patients with and without a value observed for the confounder of interest) with p variables (e.g., patient or disease characteristics) whether their mean vectors (p-dimensional vectors of means) are significantly different. To that end, the T^2 statistic is derived by a vector of covariate means (i.e, one element for each covariate) [6], which are represented as x bar 1 (vector of covariate means in group 1) and x bar 2 (vector of covariate means in group 2) in the equation [7] below (with n1 and n2 representing the sample sizes of group 1 and 2, respectively, and Sp representing the pooled variance-covariance matrix).

$$T^2 = (\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_{\mathbf{2}})^T \{\mathbf{S}_p(\frac{1}{n_1} + \frac{1}{n_2})\}^{-1}(\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_{\mathbf{2}})$$

To implement Hotelling's test in the `smdi` package, we built a wrapper around the `Hotelling::hotelling.test()` function [5].

- `smdi_little()`: Little's test, as opposed to Hotelling's test, performs a single global test to evaluate the missing completely at random (MCAR) assumption [8]. It was developed since with an approach like Hotelling's test, there can be concerns regarding multiplicity due to the multiple testing of each variable. Little's test first identifies a set of subgroups that share the same missing data patterns. Across these missing data patterns, it then tests for mean differences on every variable by comparing the observed means versus expected population means which are estimated using a maximum likelihood (ML) expectation-maximization (EM) algorithm as implemented by the `norm::prelim.norm()` and `norm::em.norm()` functions in `R`.

The test statistic (d^2) is derived by computing the sum of squared standardized differences between the subgroup means and the expected population means and weighted by both the estimated variance-covariance matrix and the respective subgroup sizes [9,10].

$$d^2 = \sum_{j=1}^{J} n_j(\hat{\mu}_j - \hat{\mu}j^{(\mathrm{ML})})^T \hat{\Sigma}_j^{-1}(\hat{\mu}_j - \hat{\mu}j^{(\mathrm{ML})})$$

ygedyg. To implement Little's test in the **`smdi`** package, we built a wrapper around the **`naniar::mcar_test()`** function [11].

For both `smdi_hotelling()` and `smdi_little()`, a high test statistic and a low p-value would indicate differences in the groups compared. A limitation of both Hotelling's and Little's test is that they assume continuous variables following multivariate normality. Real-world data, however, are often of binary nature and to consider categorical data in the computation of these tests, categorical data are one-hot encoded to binary dummy variables before performing each test.

- `smdi_asmd()`:

References

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

2 Moreno-Betancur M, Lee KJ, Leacy FP, *et al.* Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. *American Journal of Epidemiology.* 2018;187:2705–15.

3 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

4 Hotelling H. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics.* 1931;2:360–78.

5 Curran J, Hersh T. Hotelling: Hotelling's t^2 test and variants. Published Online First: 2021.

6 Multivariate analysis of means for two groups. https://rpubs.com/juanhklopper/multivariate_comparison_of_means_of_two_groups (accessed 19 November 2023)

7 Applied multivariate statistical analysis (PennState). https://online.stat.psu.edu/stat505/lesson/7/7.1/7.1.15 (accessed 19 November 2023)

8 Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association.* 1988;83:1198–202.

9 Little's missing completely at random (MCAR) test. https://search.r-project.org/CRAN/refmans/misty/html/na.test.html (accessed 19 November 2023)

10 Enders CK. *Applied missing data analysis.* Guilford Publications 2022.

11 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