September 15, 2023

Editorial Board

JAMIA open

Subject**:** Submission of Manuscript **"** **smdi: An R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies"**

Dear Editorial Board Members,

I am writing to submit our manuscript titled "*smdi: An R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies*" for consideration as an application note in JAMIA Open.

Electronic health records (EHR) are increasingly linked to administrative claims databases to conduct real-world evidence (RWE) studies complementing evidence coming from randomized controlled trials. Due to their detailed capture of clinical parameters, EHR database linkages can significantly improve the ability to control for confounding when estimating treatment effects in RWE studies. However, such confounders are often just partially observed in EHR which is a pervasive challenge in the statistical analysis of the data. Many analytic methods have been proposed to overcome this issue (e.g., inverse probability weighting and multiple imputation), but their appropriateness depends on assumptions based on the underlying missingness mechanism. However, the principled characterization of potentially underlying missingness processes has received only little attention and is rarely performed in practice to underpin analytic decisions made.

Backed by recent large-scale simulation results, we developed the smdi (structural missing data investigations) R package that readily implements principled diagnostics on partially observed covariate data in real-world databases. In particular, the package streamlines multiple diagnostics which can provide useful information on how to characterize the underlying missingness patterns and mechanisms. This includes 1) assessing differences in observed patient distributions between patients with and without an observed value for the partially observed covariate, 2) evaluating the ability to predict missingness based on observed data using machine learning and 3) checking if the missingness is associated with the study outcome. Our simulations suggested that the combination the assessed diagnostics successfully identified patterns that matched assumptions for a set of simulated missing data structures.

The submitted manuscript will give your readership a systematic and intuitive introduction on how this R package can be applied in the design and analysis of their research studies and provides instructions and examples on how the resulting output should be interpreted. We further provide comprehensive documentation and vignettes on the package’s associated website (<https://janickweberpals.gitlab-pages.partners.org/smdi>).

We believe that our manuscript will be of great interest to your readership and will facilitate a systematic and thorough implementation of reproducible analytic workflows in studies with partially observed data.

Thank you for considering our submission. We look forward to your positive response.

Yours sincerely,

Janick Weberpals, RPh, PhD