Multiple imputation for multilevel data with continuous and binary variables

Vincent Audigier1, Ian R. White2;3, Shahab Jolani4, Thomas P. A. Debray5;6, Matteo Quartagno7, James Carpenter7, Stef van Buuren8 and Matthieu Resche-Rigon9;10;11

Individual participant data (IPD) meta-analysis is often considered to be the gold standard method for systematic reviews. The aim is to consider several studies, sharing the same outcome, to obtain better inference than could be obtained from any one study. However, studies typically differ in their data collection and availability of confounders typically varies. Consequently, by merging the studies, systematically missing data, i.e. missing for all individuals in a study, could be introduced. In addition, missing data (called sporadically missing) can occur within each study.

Multiple imputation (MI) is a common strategy to overcome the missing data issue. The imputation model used can be an explicit joint model (JM), specifying the distribution of all variables, or it can be defined only by conditional densities (fully conditional specification, FCS). The choice of the imputation model is a task crucial but difficult a priori.

We investigate MI methods to overcome systematically and sporadically missing data in a multilevel setting, such as meta-analysis, in the context of mixed data (continuous and binary). The methods compared are JM imputation of clustered data proposed by Quartagno and Carpenter (2016), FCS using generalized mixed models proposed by Jolani et al. (2015), FCS using a two-stage meta-analysis estimation procedure (Resche-Rigon and White, 2016).

First, methods are compared from a methodological point of view and through a simulation study. The study highlights the benefit to use such methods compared to reference multiple imputation methods for multilevel missing data. However, this work also shows that performances need to be nuanced according to the missing data pattern, the multilevel structure and the type of missing variables. Then, the MI methods are applied to an IPD meta-analysis in cardiovascular disease consisting of 28 observational cohorts in which systematically missing and sporadically missing data occur. Finally, practical recommendations are provided.