A framework for individual participant data meta-analysis in the presence of missing data

Individual Participant Data Meta-Analysis with systematically missing predictors

TPA Debray[1], H Koffijberg[1], S Jolani[3], Y Vergouwe[2], D Nieboer[2], EW Steyerberg[2], S Van Buuren[3], KGM Moons[1]

Further collaboration planned with I White and M Resche-Rigon.
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
[2]Center for Medical Decision Sciences, Eramsus Medical Center Rotterdam, The Netherlands
[3]Department of Methodology and Statistics, Utrecht University, The Netherlands

UMC Utrecht

Introduction

Individual participant data meta-analyses (IPD-MA) are increasingly used for developing multivariable risk prediction models. Unfortunately, some predictors may not have been measured in each study and are systematically missing in the IPD-MA. As a consequence, researchers often choose to exclude entire studies with one or more missing predictors from the IPD-MA. Alternatively, systematically missing predictors are ignored during model development. It may be clear that this approach is undesirable as available evidence is not optimally used, and certainly if the missing predictors are known to be important.

Approaches

- Complete Case Analysis (CCA)
  - excludes studies where important predictors have not been measured
  - assumes that the occurrence of systematic missingness in predictors is MCAR on the study level.
- Full Predictor analysis (FPA)
  - simply discard systematically missing predictors during model development
  - no assumptions are made about the missing data mechanisms
- Traditional Multiple Imputation (TMI)
  - stack all study IPD and treat them as a single dataset during imputation
  - missingness mechanisms depend on the observed data only (MAR)
  - does not account for clustering of subjects within studies, and assumes a common covariance structure for all IPD.
- Stratified Multiple Imputation (SMI)
  - Extend the imputation model with a clustering term
  - Stratified intercept term
  - Random slope(s) (**To be implemented**)

Case Study

- Diagnosis of Deep Vein Thrombosis
- Develop a logistic regression model with a predefined set of 8 predictors
- IPD meta-analysis of 13 studies ($N = 10\,002$, with 1 864 events)
- 10 fully measured predictors
- 5 systematically missing predictors (results D-dimer test, presence leg trauma, family history of thrombofilia, oral contraceptive use, history of previous DVT)
- External validation of the model in an independent study with no missing data

Main findings

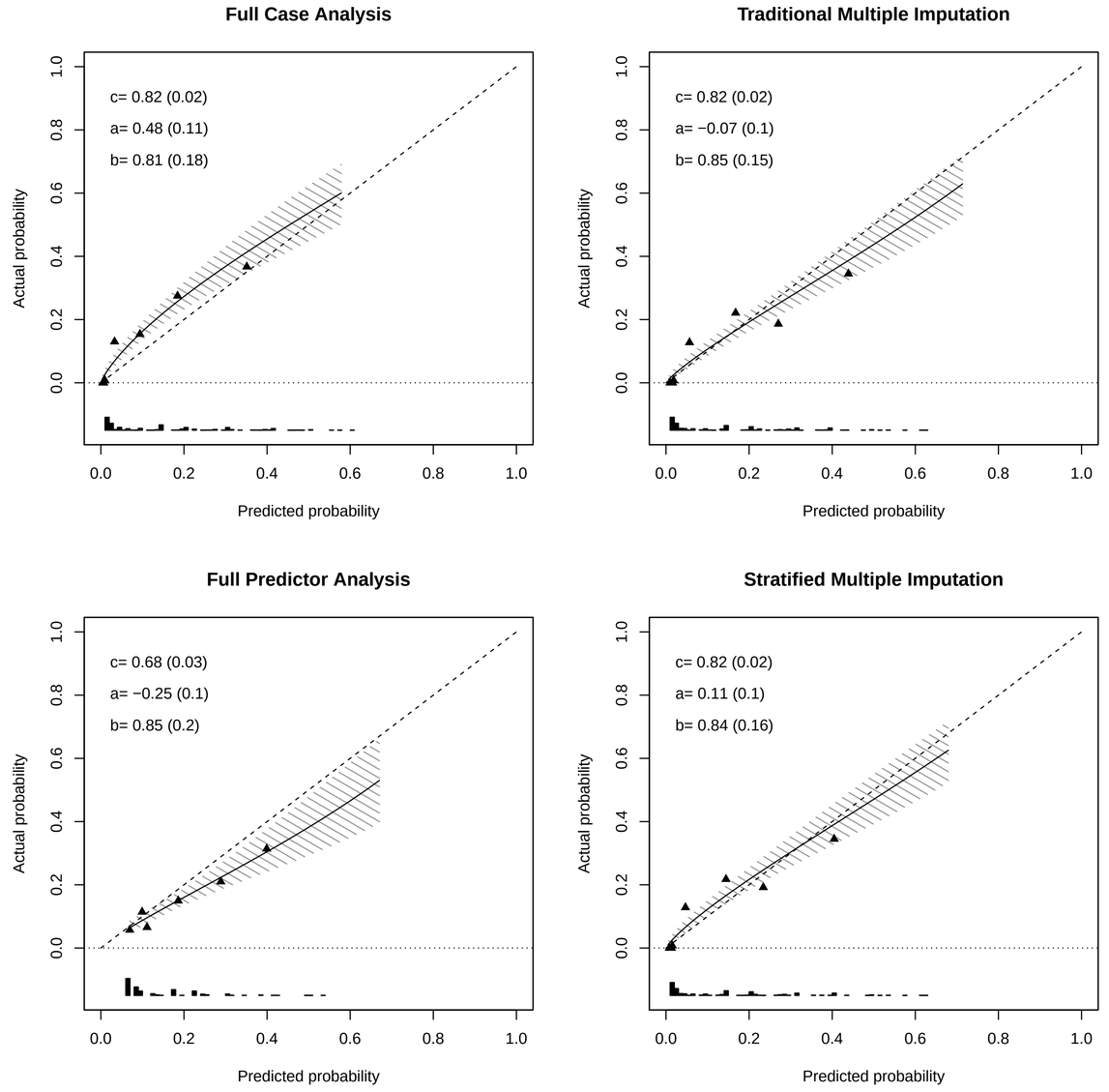
- TMI generally leads to sufficient model performance
- SMI theoretically more sound, but yields similar performance (random slopes models still to be evaluated)
- CCA substantially hampers model development when few studies remain in the analysis
- FPA leads to poorest model performance, particularly when important predictors are ignored during mode development

Contact

Thomas Debray, MSc
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
Tel: +31 (0) 88 75 68640
Email: t.debray@umcutrecht.nl

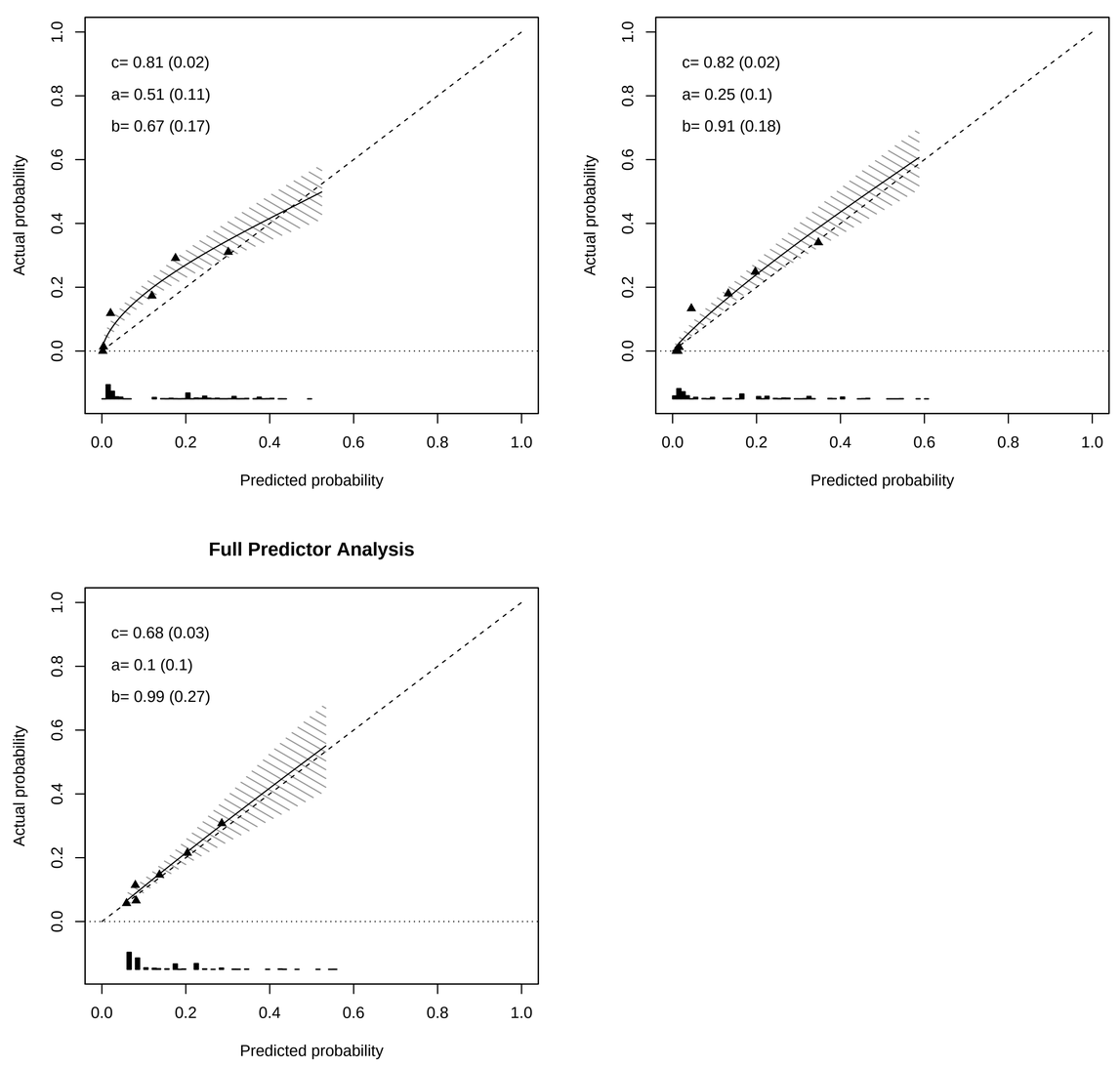
Results Scenario 1

12 studies available for model development ($N = 8\,974$)



Results Scenario 2

6 studies available for model development ($N = 4\,466$), with 5 studies affected by systematic missingness



This work was supported by Netherlands Organisation for Scientific Research (NWO) (9120.8004, 918.10.615 and 916.11.126)