UMC Utrecht
Julius Center

Multiple imputation of systematically missing predictors in an individual participant data meta-analysis

Debray TPA, Jolani S, Koffijberg H, van Buuren S, Moons KGM

Individual Participant Data meta-analysis

- **Intervention research**
  - Assessment of treatment efficacy
  - Effect modification & subgroup analysis
- **Diagnostic research**
  - Diagnostic test evaluation (e.g. accuracy: DTA)
  - Development & validation of prediction models
- **Prognostic research**
  - Prognostic factor research
  - Development & validation of prediction models

By using datasets from multiple studies, it becomes possible to **address between-study heterogeneity** and **investigate generalizability** across different study populations

IPD meta-analysis and missing data

- Common to **impute datasets separately** due to potential for between-study heterogeneity
  - differences in outcome prevalence/incidence
  - differences in associations (e.g. treatment effect)
- Separate imputation is problematic when some (important) variables are not measured in each individual dataset
  - Exclusion of studies with missing variables
  - Omission of missing variables from the analyses
  - Implementation of (naïve) imputation strategies

Advanced imputation strategies are needed to account for systematically missing data in an IPD-MA

Imputation of continuous systematically missing variables

Previously, *Resche-Rigon* et al. developed a multiple imputation approach that[1]:

- Is based on MICE (conditional imputation model)
- Assumes missing at random (MAR)
- Adopts a linear mixed effect model with random intercept term and slopes

[1] Resche-Rigon M et al. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Stat Med. 2013 Dec 10;32(28):4890-905.

Imputation of non-continuous systematically missing variables

Approach of Resche-Rigon et al becomes problematic

- Non-continuous data: binary, categorical, count, …
- Estimation of mixed effects models more complex
- Technical issues arise around estimation of covariance parameters
- Need for alternative assumptions in imputation model

Imputation of continuous and non-continuous systematically missing variables

- MICE procedure (assuming MAR)
- Generalized linear mixed effect model with
  - Fixed effects parameters
  - Between-study covariance parameters (modeled by an inverse Wishart distribution)
  - Dispersion parameter(s) (only for imputation of continuous predictors)
- Diffuse prior distributions

Empirical example

Diagnosis of deep vein thrombosis (DVT) in patients with a suspected DVT

- IPD meta-analysis of **13 studies** (N=10,002)
- **Methods**: investigate between-study heterogeneity in a predefined set of 8 predictor variables (taken from an existing model developed by *Oudega*)
- **Aim**: assess whether the predictor variables can reliably be used in a novel prediction model

(if there is much heterogeneity, model performance will be inconsistent across study populations)

Empirical example

Diagnosis of deep vein thrombosis (DVT) in patients with a suspected DVT

- 11 predictors measured in all studies
  - Presence of malignancy (*malign*)
  - …
- 4 (binary) predictors systematically missing
  - Results D-dimer test (*ddimd*)
    missing in 5 studies
  - Family history of thrombofilia (*notraum*)
    missing in 7 studies
  - Leg trauma presence
    missing in 6 studies
  - Use of oral contraceptives
    missing in 8 studies

Empirical example

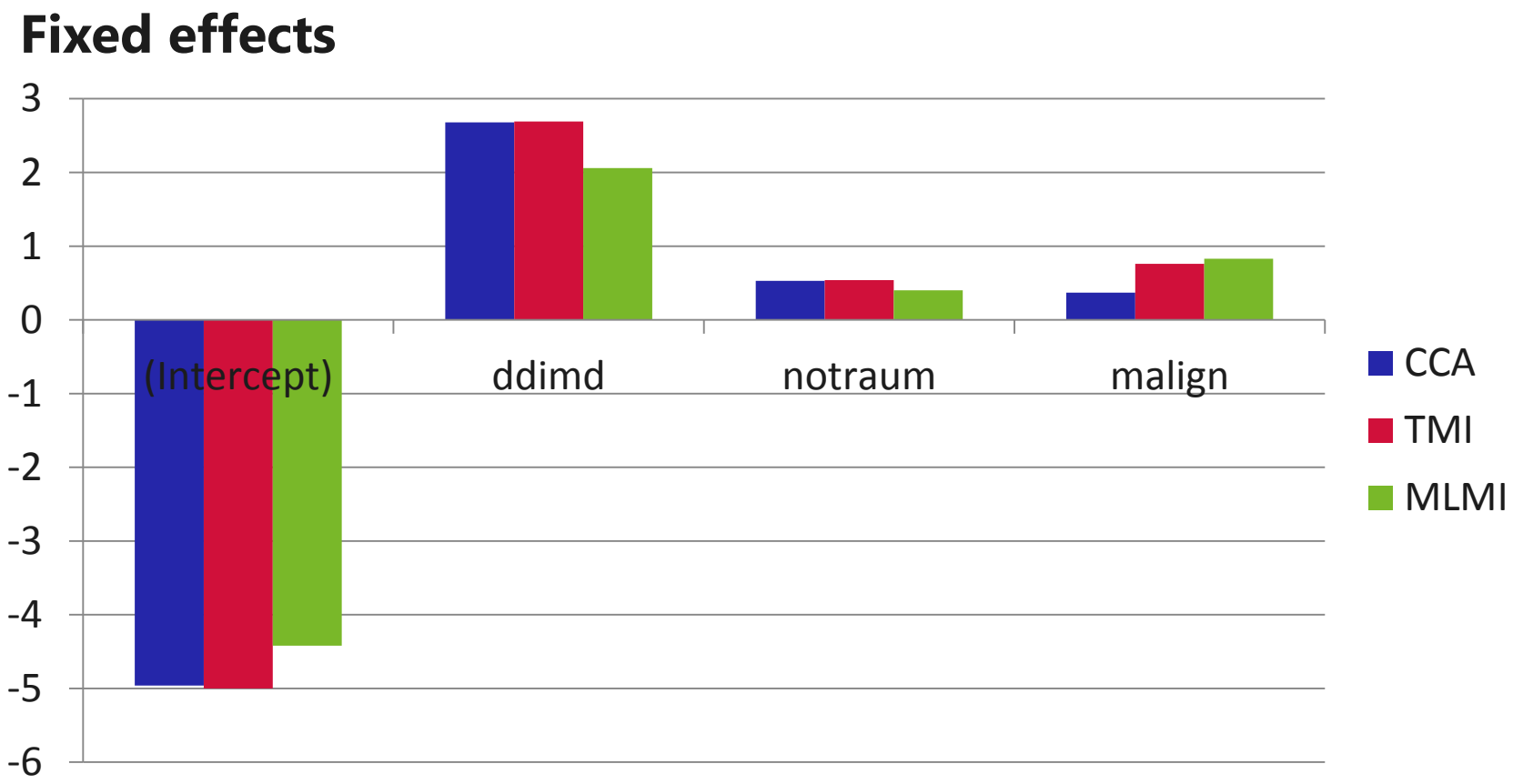
Methods for imputation

- Complete case analysis (**CCA**)
  exclude studies with missing predictor
  reduces the IPD-MA from 13 to 4 studies
- Traditional multiple imputation (**TMI**)
  imputation model ignoring between-study heterogeneity
- Multilevel multiple imputation (**MLMI**)
  imputation model accounting for between-study heterogeneity

Methods for data analysis

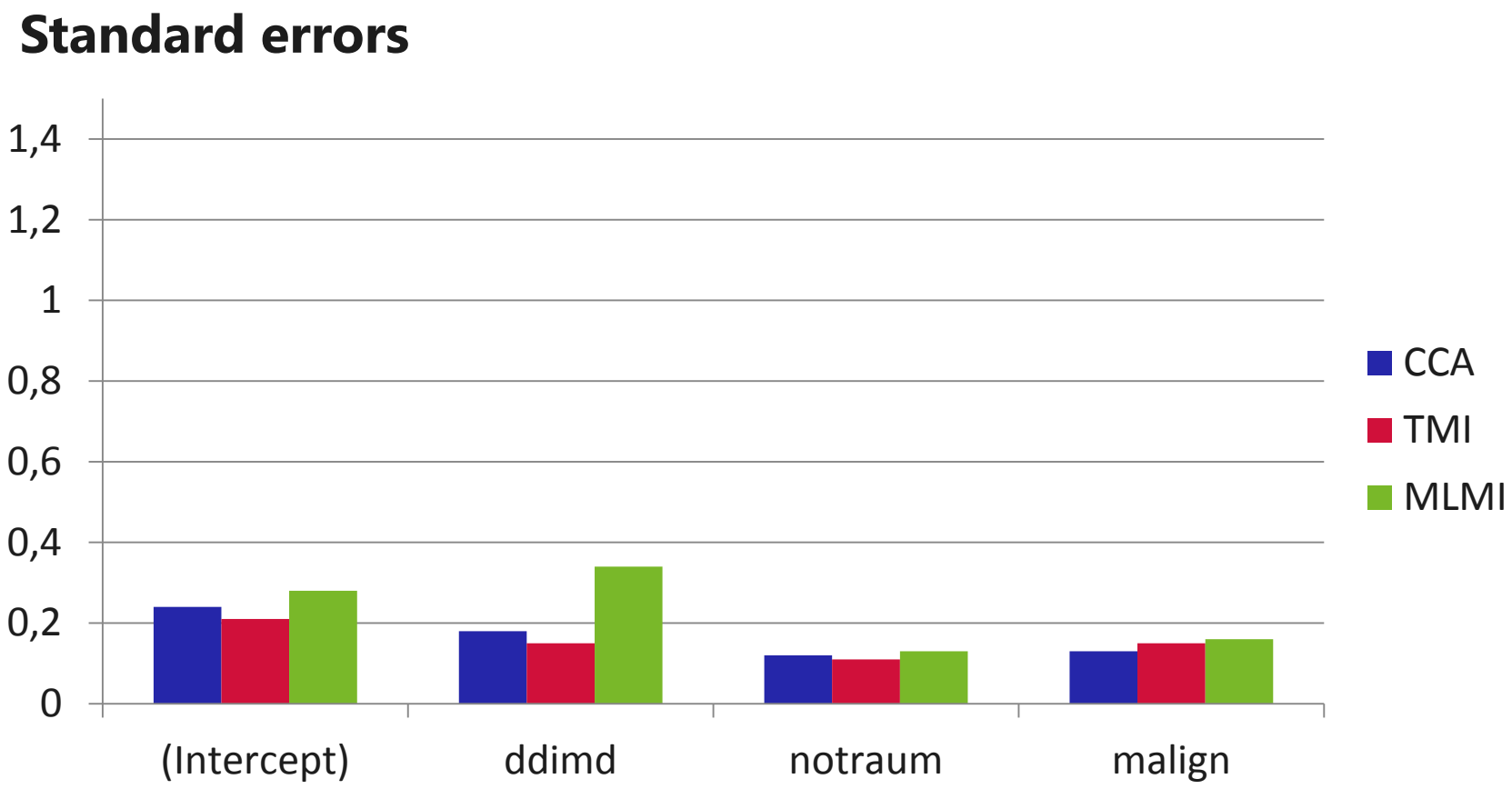
- Estimation of mixed effect model with joint random effects on all 8 predictor variables (+ intercept term)

Empirical example results



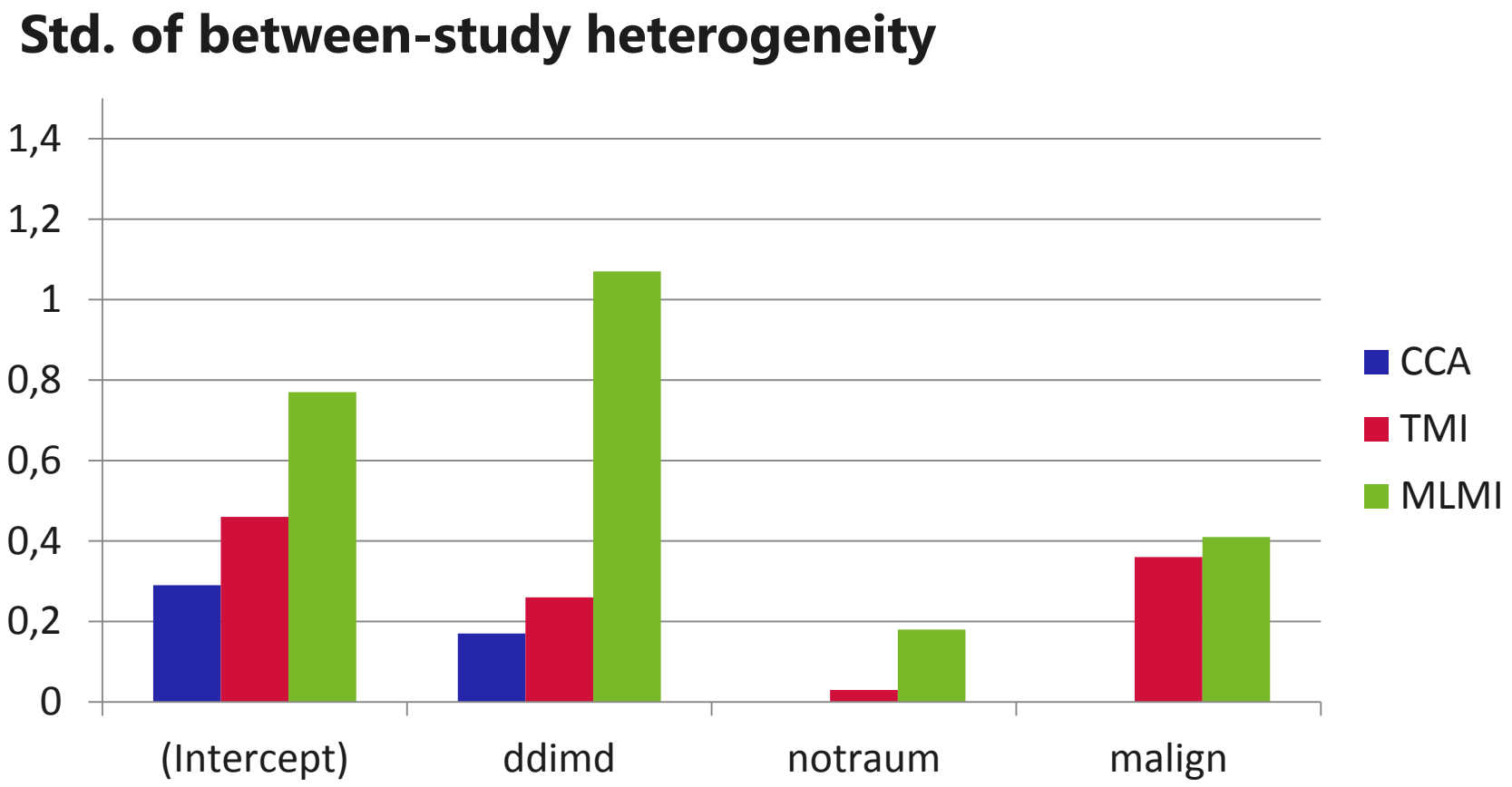
CCA = complete case analysis
TMI = traditional multiple imputation
MLMI = multilevel multiple imputation

Empirical example results



CCA = complete case analysis
TMI = traditional multiple imputation
MLMI = multilevel multiple imputation

Empirical example results



CCA = complete case analysis
TMI = traditional multiple imputation
MLMI = multilevel multiple imputation

Simulation study

- Based on DVT case study, but using 2 predictors that were measured in all studies
- Introduction of systematically missing predictors according to MCAR

Results (not shown)

- Fixed effect estimates (predictor effects)
  - Similar estimates for all methods
  - Problematic coverage for TMI and CCA
- Between-study heterogeneity estimates
  - Too low when using CCA or TMI
  - Sometimes too large when using MLMI

Discussion

- **CCA**
  - Underestimates actual degree of heterogeneity
  - Problematic when MCAR is not justified
  - Problematic when multiple variables are missing, and almost all studies need to be excluded
- **TMI**
  - Underestimates actual degree of heterogeneity
- **MLMI**
  - Optimal coverage (predictor effects)
  - Lowest bias (between-study heterogeneity)
  - Possible issues: convergence & model complexity

Take home message

Use of multilevel imputation recommended to properly identify between-study heterogeneity

- **Diagnosis & prognosis research**
  - Inclusion of heterogeneous predictors may degrade model generalizability and lead to inconsistent performance
  - Heterogeneity in DTA may lead to unfavorable (treatment) decisions in new study populations
- **Intervention research**
  - Heterogeneity in treatment effect (or treatment-covariate interactions) may indicate the presence of confounding, effect modification, or bias
  - Heterogeneity -> **red flag** when recommending treatments in certain populations or patients