Dealing with missing data in an Individual Participant Data Meta-Analysis

TPA Debray[1,2], S Jolani[3], A Schierenberg[1], KGM Moons[1,2]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
[2]Cochrane Netherlands, University Medical Center Utrecht, The Netherlands
[3]Department of Methodology and Statistics, Maastricht University, Maastricht, The Netherlands

UMC Utrecht

Background & Objective

- It is well known that the presence missing data may lead to substantial bias and reduced statistical power
- Multiple imputation is generally recommended to adequately propagate uncertainty arising from missing data
- Lack of guidance for dealing with missing data across multiple data sources, such as individual participant data meta-analyses (IPD-MA)

Aim: To compare several methods for imputing missing data in an IPD-MA and synthesizing the corresponding results.

Available Methods

Dealing with missing data

- *Complete case analysis*: Remove individuals with missing values.
- *Within-study imputation*: Impute each study dataset separately
- *Stratified imputation*: Stack all study datasets and impute them together. Imputation is based on generalized linear effects models where the study variable is treated as a dummy factor.
- *Hierarchical imputation*: Stack all study datasets and impute them together. Imputation is based on generalized linear *mixed* effects models where random effects are assumed for one or more coefficients.

Synthesis of data sources

- *One-stage meta-analysis*: Each completed version of the IPD-MA dataset is analysed using a single statistical model that accounts for potential between-study heterogeneity.
- *Two-stage meta-analysis*: A separate model is first fitted in each completed study dataset. Afterwards,
  - Apply meta-analysis for each completed version of the IPD-MA, and combine the meta-analysis results using Rubin's rules (MA-RR).
  - Combine study-specific estimates using Rubin's rules and then meta-analyse the combined estimates across studies (RR-MA).

Method		Imputation	Meta-Analysis	Order of Pooling
Acronym	Symbol	Description	Description	
CO	△	Complete case analysis	One-stage	-
HO†	□	Hierarchical imputation	One-stage	-
HO‡	○	Hierarchical imputation	One-stage	-
SO	◇	Stratified imputation	One-stage	-
WO	▽	Within-study imputation	One-stage	-
$\text{HT}^{\dagger}_{\text{marr}}$	+	Hierarchical imputation	Two-stage	MA-RR
$\text{HT}^{\dagger}_{\text{rrma}}$	×	Hierarchical imputation	Two-stage	RR-MA
WT_{marr}	⊞	Within-study imputation	Two-stage	MA-RR
WT_{rrma}	⊠	Within-study imputation	Two-stage	RR-MA

† Imputation allows for heteroscedastic within-study error variances is achieved using a fully Bayesian Gibbs sampler
‡ Imputation assumes homoscedastic within-study error variances and is achieved using large sample approximations.

Simulation study

Generation of IPD-MA with 10 studies of 250 participants each

- Binary outcome
- Two continuous covariates with varying mean, covariance and covariate-outcome association across studies
- Missing values for one or more covariates according to MAR

Analysis: Five imputations were created for each incomplete data set. All meta-analysis models allowed for joint random effects on the intercept term and regression coefficients.

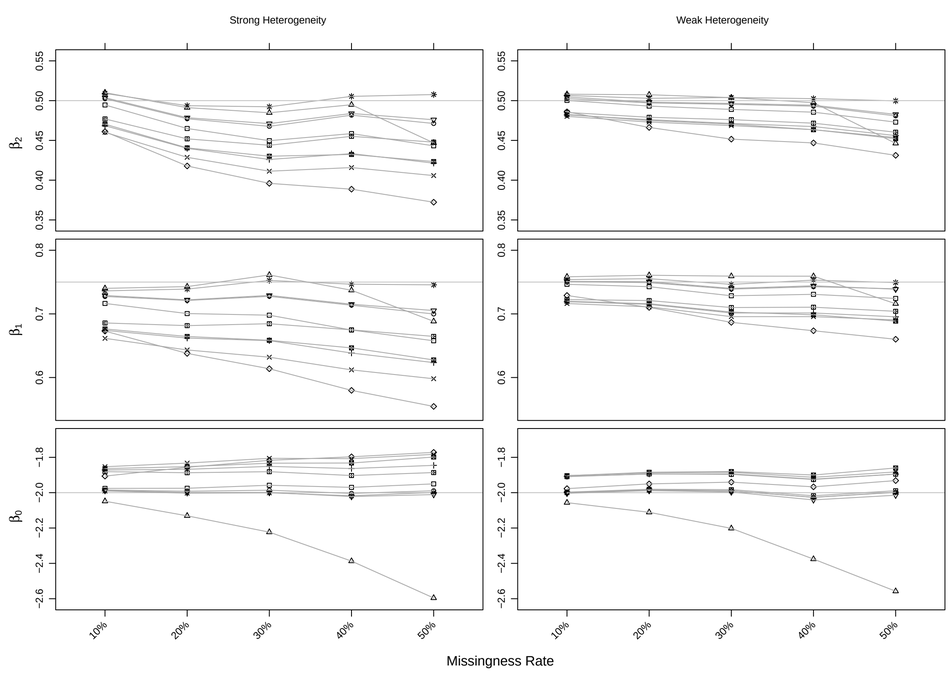

Figure : Results for fixed effects estimates in the simulation study.

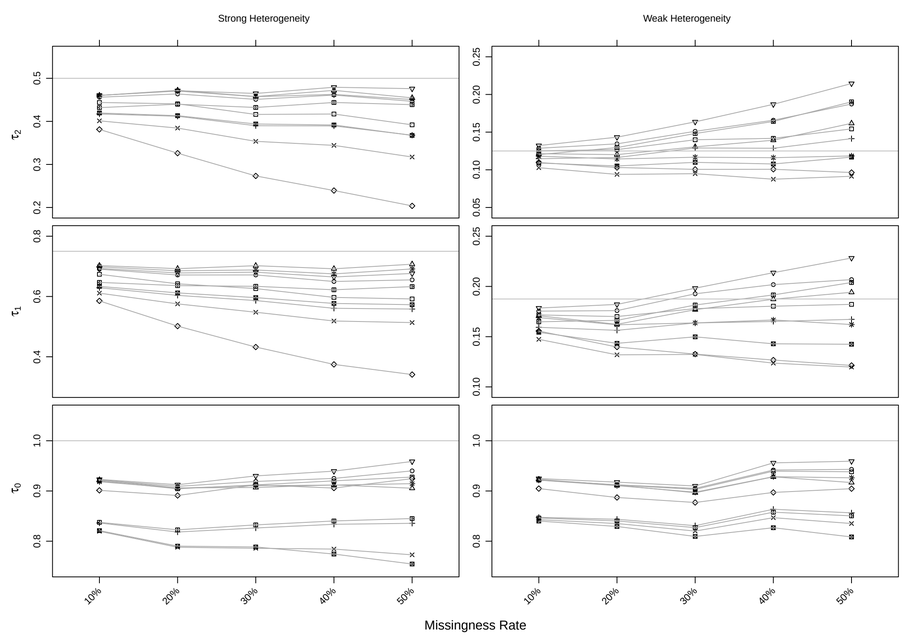

Figure : Results for estimates of between-study variability of the regression coefficients in the simulation study.

Case Study

- **Data**: IPD from 7 cross-sectional studies examining the diagnostic accuracy of the inflammation marker C-reactive protein (CRP) in capillary blood
- **Primary outcome**: community acquired pneumonia (CAP) in primary care, determined by chest radiography.
- **Covariate of interest**: multivariable coefficient of CRP, modeled as $\ln(1 + \text{CRP})$.
- **Adjustment covariates**: age, sputum production, dyspnoea, and current temperature
- **Introduction of missing data**: *Random* missing values in each study dataset for CRP and temperature measurements (subjects with one or more missing values: 12 – 34% per dataset).

One-stage IPD-MA	Ref.	CO	HO†	HO‡	SO	WO
β	0.90	0.88	0.93	0.86	0.91	0.89
SE(β)	0.07	0.06	0.06	0.08	0.05	0.09
τ_{β}	0.10	0.07	0.06	0.12	0.05	0.16
Two-stage IPD-MA	**Ref.**	**CT**	$\text{HT}^{\dagger}_{\text{marr}}$	$\text{HT}^{\dagger}_{\text{rrma}}$	WT_{marr}	WT_{rrma}
β	0.89	0.84	0.91	0.91	0.85	0.87
SE(β)	0.08	0.07	0.06	0.06	0.11	0.07
τ_{β}	0.14	0.10	0.08	0.07	0.22	0.12

Table : Multivariable regression coefficient for transformed CRP values.

Ref = Results obtained by analyzing the original data, *before* missing values are introduced.

Conclusions

- Use of complete case analysis or stratified imputation (by extending the imputation model with a dummy variable indicating study membership) is flawed and may lead to substantial bias.
- Hierarchical imputation and within-study imputation perform similarly, although the former tends to yield more accurate results.
- When the amount of studies and participants per study is sufficiently large, within-study imputation followed by two-stage meta-analysis may be preferred to avoid difficult modeling choices and speed up the estimation procedure.
- Two-stage IPD-MA that are based on imputed datasets should first apply meta-analysis for each completed version of the IPD-MA, and then combine the meta-analysis results using Rubin's rules (MA-RR).

Contact

Please do not hesitate to get in touch with us during the ISCB!

Thomas Debray, PhD
Assistant Professor
T.Debray@umcutrecht.nl

Shahab Jolani, PhD
Assistant Professor
s.jolani@maastrichtuniversity.nl

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