Systematically missing covariates in IPD meta-analysis

Comparing different statistical approaches using large cardiovascular trials

Robert Thiesmeier

Karolinska Institutet, Stockholm, Sweden

Royal Statistical Society International Conference, Edinburgh 2025

Missing cardiac biomarker values across trials

	Trial 1 (N=7,249)	Trial 2 (N=6,157)	Trial 3 (N=11,063)	Trial 4 (N=6,522)	Trial 5 (N=7,067)
Age, y	63.8 (6.8)	65.0 (8.6)	63.7 (8.3)	64.3 (9.5)	65.4 (8.3)
Female, n (%)	2,731 (37.7)	2,039 (33.1)	3,907 (35.3)	1,669 (25.6)	1,717 (24.3)
BMI $\geq$ 30, n (%)	4,342 (59.9)	3,443 (55.9)	9,604 (86.8)	2,034 (31.2)	2,240 (31.7)
NT-proBNP, pg/mL	75.2	139.5	92.0	**NA**	**NA**

Statistical approaches for missing data in multi-site studies

1. **Partially adjusted analysis**
   - Exclude the missing variable from the analysis model at specific sites
   - Risk of bias due to lack of adjustment for key variables
   - Avoids loss of participants and exclusion of study sites
2. **Complete case analysis**
   - Exclude studies with 100% missing data
   - Shown to be *unbiased* if missing mechanism is MCAR
   - Risk of bias in heterogenous settings and loss of participants
3. **Bivariate meta-analysis**
4. **Two-stage multiple imputation**

Bivariate Meta-Analysis

- A bivariate meta-analysis for systematically missing covariates allows joint modelling of partially (θ_{iP}) and fully (θ_{iF}) adjusted estimates
- Borrow strength at the study level by using the observed association between fully and partially adjusted estimates in studies where both are available
- It allows all studies to contribute toward the summary estimate of the fully adjusted effect (θ_F)

Bivariate meta-analysis

- Partially (θ_{iP}) and (where possible) fully (θ_{iF}) adjusted estimates are obtained from each study, together with their standard errors
- The bivariate model can be written as:

$$\begin{pmatrix} \hat{\theta}_{iF} \\ \hat{\theta}_{iP} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta_{iF} \\ \theta_{iP} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$
$$\begin{pmatrix} \hat{\theta}_{iF} \\ \hat{\theta}_{iP} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta_{F} \\ \theta_{P} \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \kappa\tau_1\tau_2 \\ \kappa\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right) \quad (1)$$

- The ρ and κ represent the within and between-study correlation, respectively
- The within-study correlation (ρ) can be estimated using a non-parametric bootstrap approach

Two-stage multiple imputation

- Multiple imputation allows imputation of estimates on the participant level
- We followed a two-stage imputation process for multi-site studies (Resche-Rigon et al. 2018) with quantile regression (Thiesmeier et al. 2024)
- Quantile regression can be a flexible imputation method as it makes no distributional assumptions

Two-stage multiple imputation: Quantile regression

- In a study with collected data on z_i, estimate p-quantile regression model for z_i conditionally on a set of predictors $\mathbf{w}_i$:

$$\hat{Q}_{z_i|\mathbf{w}_i}(p) = \mathbf{w}_i\mathbf{f}(p) \quad p \in \{0.01, 0.02, \ldots, 0.99\} \tag{2}$$

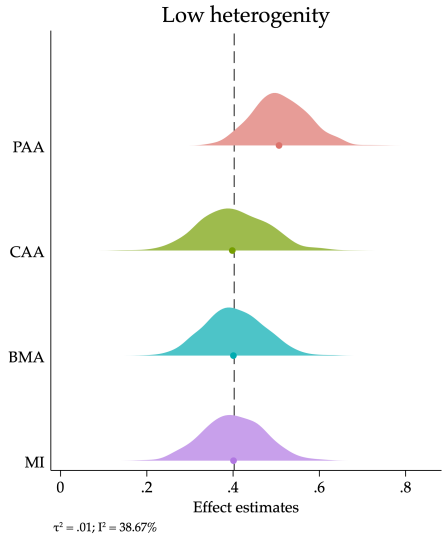
- In the study with missing data, draw a random value U_i from a continuous uniform distribution $\mathcal{U}(0, 1)$

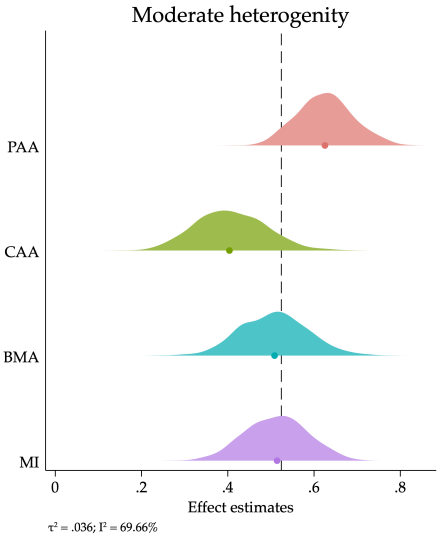
- Compute the weighted average of the F and $F + 1$ conditional predicted quantiles and assign:

$$z_i^{(m)} = (1 - \text{mod}) \cdot \hat{Q}_{z_i|\mathbf{w}_i}(F) + \text{mod} \cdot \hat{Q}_{z_i|\mathbf{w}_i}(F + 1) \tag{3}$$

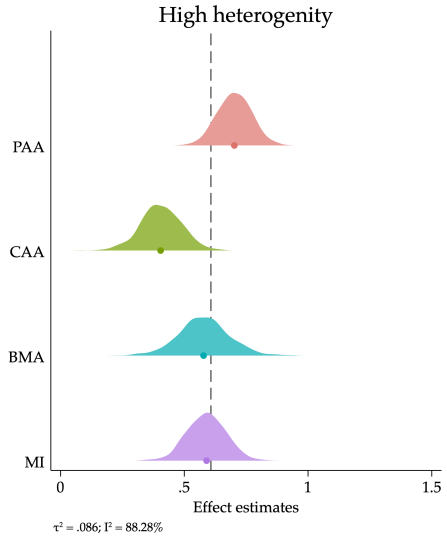
where $F = \lfloor U_i\% \rfloor$ and $\text{mod} = U_i\% - \lfloor U\% \rfloor$

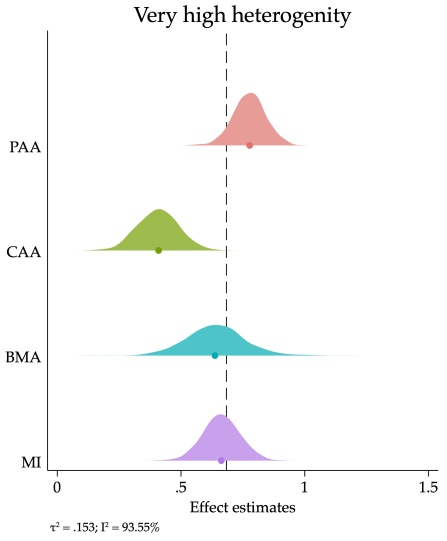
Simulations: Varying heterogeneity between studies





Simulations: Varying heterogeneity between studies



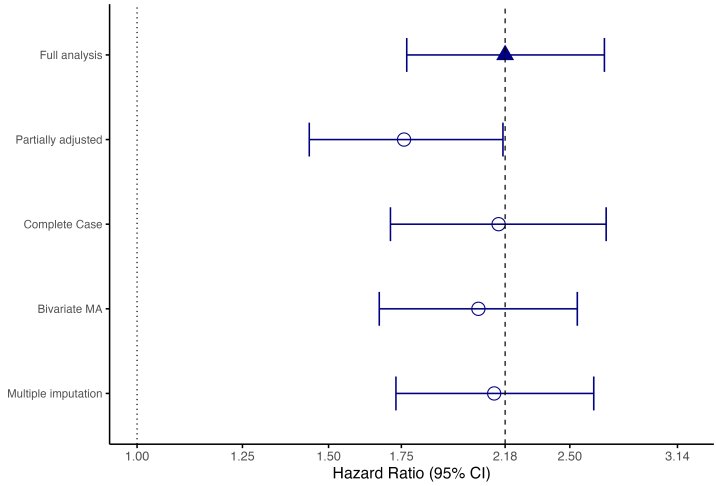


Application to large cardiovascular trials

- We pooled 3 large trials ($N = 24,469$) to assess the prognostic performance of BMI on hospitalization due to heart failure (HHF) including NT-proBNP as a strong confounding variable
- Two trials had observed data on NT-proBNP, whereas one trial had systematically missing data on NT-proBNP
- We applied all four approaches to account for the systematically missing NT-proBNP values

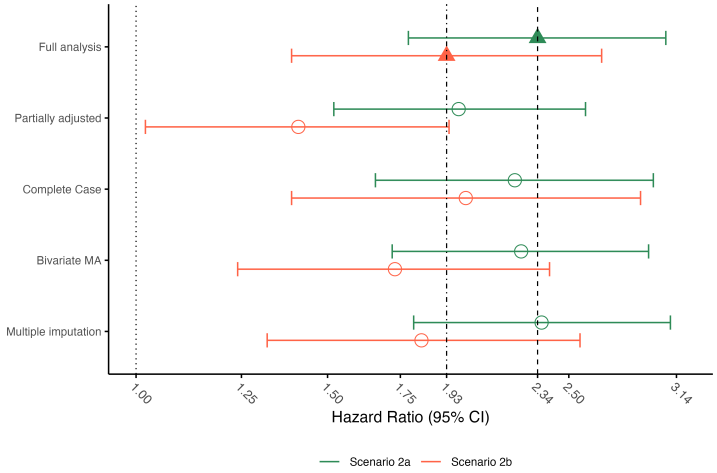
Clinically homogenous scenario

- Similar inclusion/exclusion criteria
- All trials were harmonized for analysis *without* modifications



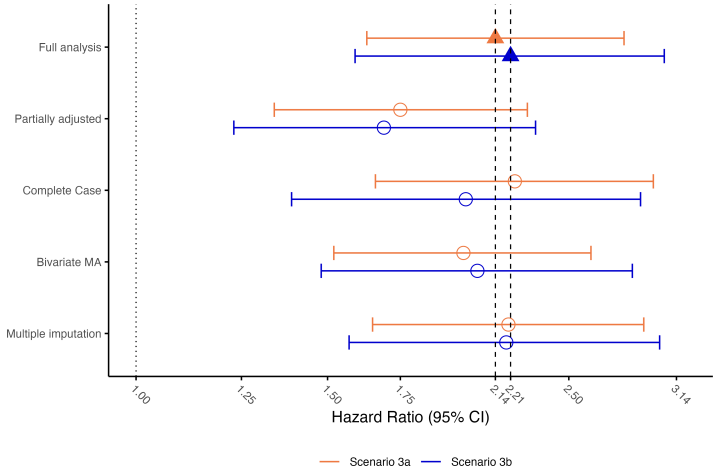
Stratified scenario

- Stratification by history of heart failure
- **Scenario 2a**: Participants *with no* prior HF
- **Scenario 2b**: Participants *with* prior HF



Clinically heterogenous scenario

- Varying artificial inclusion/exclusion criteria
- **Scenario 3a**: 2 trials included participants *without* prior HF and *complete* NT-proBNP data whereas 1 trial included participants *with* prior HF
- **Scenario 3b**: reverse criteria to 3a



Final notes

- **Partially adjusted analysis** (omitting missing variable from the analysis) can lead to strong bias
- **Complete case analysis** can be valid if studies are more homogenous and loss of participants is not a major concern
- **Bivariate meta-analysis** performs well but only recovers a single study estimate for the sites with missing data
- **Multiple imputation** has the flexibility to model the missing data (imputation model) and can thus accommodate more complex scenarios and allow for downstream analyses

Thank you for listening

robert.thiesmeier@ki.se

https://github.com/robertthiesmeier

Acknowledgements

Nicola Orsini, Matteo Bottai (Karolinska Institutet)
Andrea Bellavia, Sabina Murphy & the TIMI Study Group (Harvard University)

Karolinska Institutet

HARVARD MEDICAL SCHOOL

References

1. Thiesmeier R, Bottai M, Orsini N. Imputing Missing Values with External Data: Applications for Multi-Site Settings and Federated Analyses. *The Stata Journal*, 2025

2. Thiesmeier R, Bottai M, Orsini N. Systematically Missing Data in Distributed Research Networks: Multiple Imputation when Data Cannot Be Pooled. *Journal of Statistical Computation and Simulation*, 2024

3. Resche-Rigon M & White I. Multiple Imputation by Chained Equations for Systematically and Sporadically Missing Multilevel Data. *Statistical Methods in Medical Research*, 2018

4. Jackson et al. Systematically Missing Confounders in Individual Participant Data Meta-Analysis of Observational Cohort Studies. *Statistics in Medicine*, 2009