Cross-Site Imputation for Recovering Variables without Individual Pooled Data

Robert Thiesmeier

Karolinska Institutet, Stockholm, Sweden

International Society for Pharmacoepidemiology Annual Meeting 2025

Conflict of interest

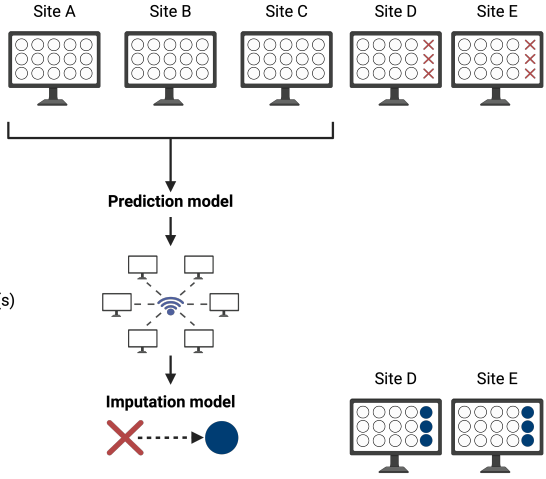
This research was supported by the National Infrastructure NEAR, supported by the Swedish Research Council [grant numbers Dnrs 2017-00639 and 2021-00178]; the National Institute for Aging (Denmark); and the National Institute of Neurological Disorders and Stroke (1R01NS131433-01) (Denmark). No relationships to disclose.

Inconsistent data across study sites

	Site A (N=136,893)	**Site B** (N=72,227)	**Site C** (N=164,687)	**Site D** (N=52,219)	**Site E** (N=43,362)
Exposure (%)	3,091 (2.3)	1,568 (2.2)	4,590 (2.9)	1,588 (3.1)	1,028 (2.5)
Confounder (%)	46,667 (34.1)	22,462 (31.1)	48,411 (29.4)	**NA**	**NA**
Outcome (%)	13,577 (9.9)	4,244 (5.9)	13,143 (8.0)	4,317 (8.3)	3,819 (8.8)

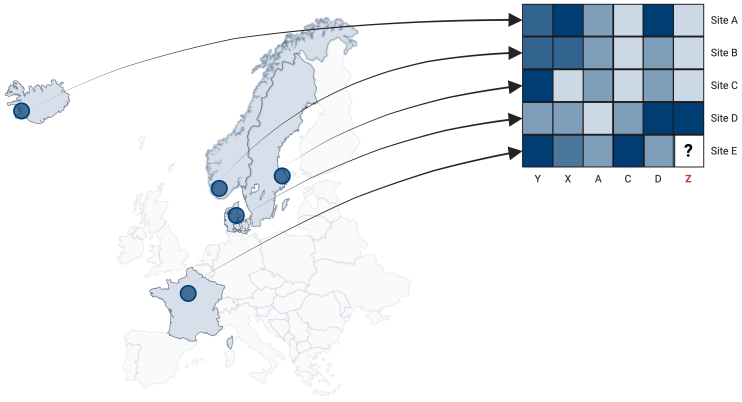
Framework for cross-site imputation

1. Identify study site(s) with observed data
2. Fit a prediction model at study site(s) with observed data on the systematically missing variable(s)
3. Transfer regression coefficients to study site(s) with systematically missing variable(s)
4. Impute systematically missing variable(s)



Identify study site(s) with observed data

Consider a multi-site study with five contributing sites and a binary variable z_i that is 100% missing at site E



Fit a prediction model at study site(s) with observed data on the systematically missing variable

Let θ_i be the probability that $z_i = 1$, given a set of predictors $\mathbf{c}_i$:

$$\theta_i|\mathbf{c}_i = P(z_i = 1|\mathbf{c}_i) \tag{1}$$

At site C, we specify a logistic regression model:

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \hat{\boldsymbol{\gamma}}_i(\mathbf{c}_i) \tag{2}$$

where $\hat{\boldsymbol{\gamma}}_i(\mathbf{c}_i)$ is the estimated linear predictor that we can send across sites

Impute the systematically missing variable

We denote $z_i^{(m)}$ as the m-th imputation of a missing value in z_i. At site E:

1. Import $\hat{\boldsymbol{\gamma}}_i(\mathbf{c}_i)$ and the related covariance structure
2. The predicted conditional probability $\hat{\theta}_i$ can be expressed as:

$$\hat{\theta}_i = \frac{e^{\hat{\boldsymbol{\gamma}}_i(\mathbf{c}_i)}}{1 + e^{\hat{\boldsymbol{\gamma}}_i(\mathbf{c}_i)}} \tag{3}$$

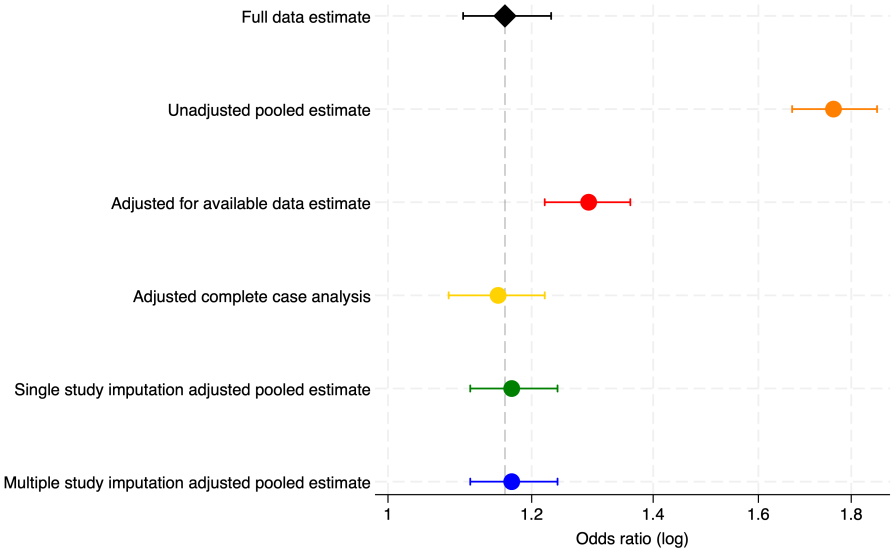
3. Draw a random value U_i from a continuous uniform distribution $\mathcal{U}(0, 1)$
4. Assign $z_i^{(m)} = 1$ if $U_i < \hat{\theta}_i$ and 0 otherwise

Maternal antidepressants and offspring neurodevelopmental disorders

- We want to study the effect of maternal antidepressant use in pregnancy on offspring risk of neurodevelopmental disorders (NDD) (ASD, ADHD, or ID)
- We need to control for a potential confounder: Parental history of psychiatric diagnosis
- Hospital 4 and 5 never recorded data on parental psychiatric history and individual data *cannot be shared data* between sites

	Hospital 1 (N=136,893)	**Hospital 2** (N=72,227)	**Hospital 3** (N=164,687)	**Hospital 4** (N=52,219)	**Hospital 5** (N=43,362)
Exposure (%)	3,091 (2.3)	1,568 (2.2)	4,590 (2.9)	1,588 (3.1)	1,028 (2.5)
Confounder (%)	46,667 (34.1)	22,462 (31.1)	48,411 (29.4)	**NA**	**NA**
Outcome (%)	13,577 (9.9)	4,244 (5.9)	13,143 (8.0)	4,317 (8.3)	3,819 (8.8)

Maternal antidepressants and offspring NDD



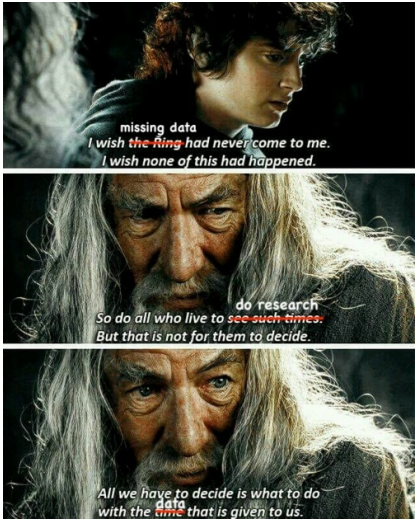
Cross-site imputation has two central assumptions

1. **Site similarity**: To what extent does the equality of causal structures between study sites impact the performance of cross-site imputation?

2. **Predictor availability**: How does variation in the availability and strength of predictors across sites affect the accuracy and bias of imputed values at sites with missing data?

Final Notes

- Multiple imputation for systematically missing values fails when individual-level data cannot be pooled
- Cross-site imputation recovers missing variables without pooling data
- The method has been implemented in Stata software and will be further improved and developed



Reference

ELSEVIER

Journal of Clinical Epidemiology

Volume 184, August 2025, 111820

Original Research

Cross-site imputation can recover missing variables in federated multicenter studies

Robert Thiesmeier [a b], Paul Madley-Dowd [c d e], Nicola Orsini [a 1], Viktor H. Ahlqvist [f g h 1]

Acknowledgements

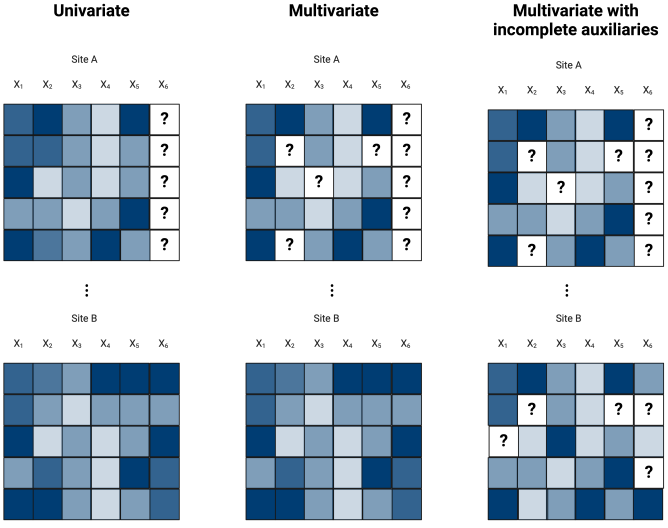
robert.thiesmeier@ki.se

https://github.com/robertthiesmeier

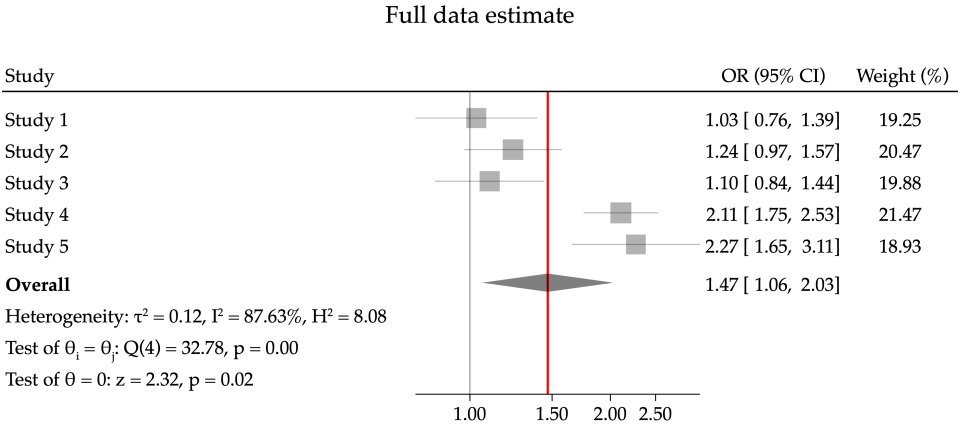
Acknowledgements

Paul Madley-Dowd (University of Bristol, UK)
Nicola Orsini (Karolinska Institutet, Sweden)
Viktor Ahlqvist (Aarhus University, Denmark)

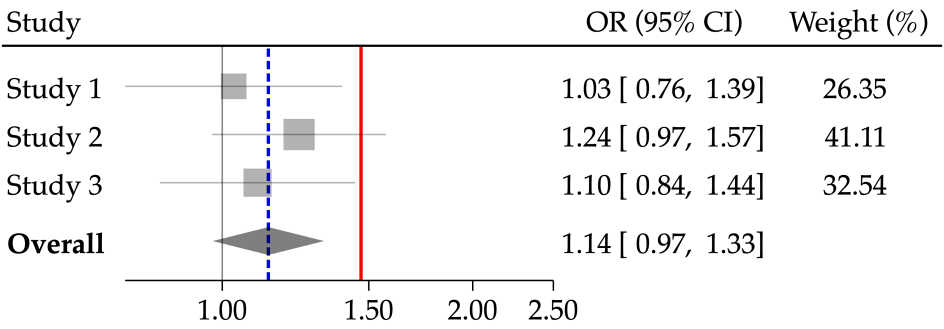
Multivariate missing data



When does a complete case analysis fail?



When does a complete case analysis fail?



Cross-site imputation to the rescue!