Multiple imputation for recovering missing values when data cannot be shared

Robert Thiesmeier

Stata Biostatistics and Epidemiology Symposium, February 20, 2025

Karolinska Institutet, Stockholm, Sweden

Overview

- Missing values in distributed data networks
- How does `mi impute from` work?
- Applied example
- Central assumptions and next steps

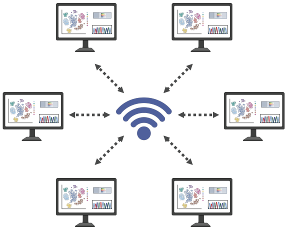
Distributed Data Networks

Medical research increasingly relies on large-scale collaborations making use of multi-site studies

- Enhances precision and can enable greater clinical granularity
- Improves generalizability, making findings applicable to diverse populations

Distributed data networks (i.e., federated analysis):

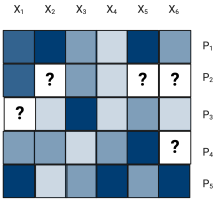
- have become the norm due to regulatory, administrative, and time constraints
- use qualitative harmonization ("common data models") and meta-analysis to avoid sharing individual-level data



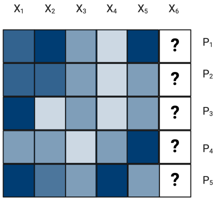
Inconsistent data across sites

- Collaborative research often faces inconsistent variable recording across sites
- Sporadically missing data: Occurs at a single site in one or more variables
- Systematically missing data: Some sites may have 100% missing data on key variables, while others have recorded them

Sporadically missing data



Systematically missing data



Available options

Current approaches include:

1. Excluding sites without the data, reducing power and generalizability (complete case analysis)
2. Ignoring missing data and meta-analyze, risking biased inferences (e.g., confounder omission, or reduced predictive accuracy)

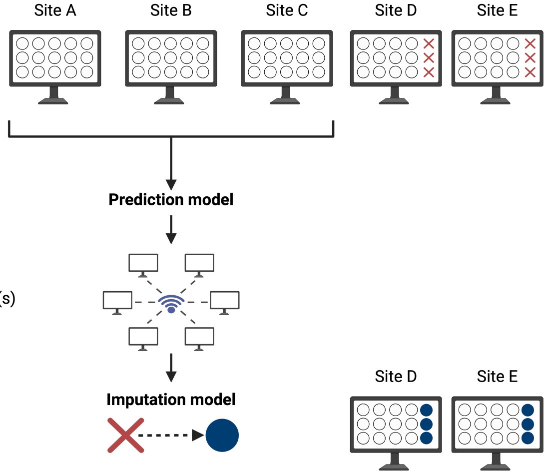
Other approaches include quantitative bias analysis or likelihood-based approaches (e.g., bivariate meta-analysis)

Can we "recover" the data instead?

- Ideally, we would like to be able to recover missing data by leveraging existing information from other sites involved in the network
- When individual-level data cannot be shared between sites, common multiple imputation strategies fail (`no observations`)
- We have proposed a "cross-site" imputation strategy that avoids the need to pool individual-level data and relies instead on sending regression coefficients across sites

Framework for multiple 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)



A new Stata command: `mi impute from`

The command `mi impute from` facilitates the imputation of variables by using external data

- Users have to specify a prediction model at sites with observed data on the systematically missing variable
- At the receiving site, `mi_impute_from_get` facilitates the convergence of shared files (.txt or .xls) to be used with `mi impute from`
- If multiple files are input, a weighted average of regression coefficients is taken

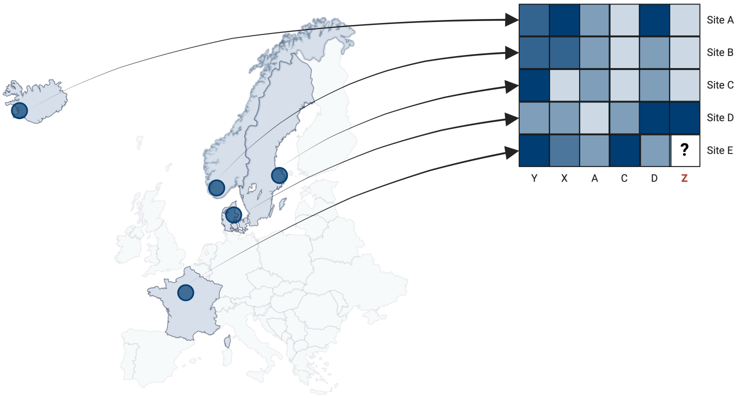
Current models that are supported include `logit`, `mlogit`, and `qreg`

Overview

- Missing values in distributed data networks
- How does `mi impute from` work?
- Applied example
- Central assumptions and next steps

Identify study site(s) with observed data

Consider a distributed data network with five contributing sites and a continuous variable z_i that is 100% missing at site E



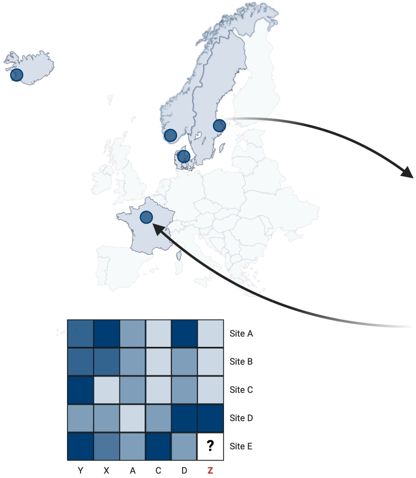
Fit a prediction model at study site(s) with observed data on the systematically missing variable

At site C, estimate p-quantile regression model for the continuous variable z_i conditionally on a set of predictors $\mathbf{w}_i$

$$\hat{Q}_{z_i|\mathbf{w}_i}(p)\& = \mathbf{w}_i\mathbf{f}(p) \quad p \in \{0.01, 0.02, \ldots, 0.99\} \tag{1}$$

If multiple studies have information on z_i, we can fit the same prediction model at multiple sites

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



- **Fit** a model using `qreg` at site with observed values on `z` (e.g., site C) using `y x a c d` as independent variables
- **Export** coefficients and their variances into a transportable file (e.g., txt) (let us call the two files `siteC_b.txt` and `siteC_v.txt`)
- **Send** files to site with missing data (i.e., site E)

Impute the systematically missing variable

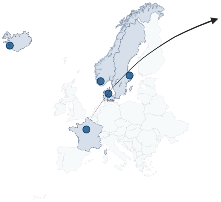
We denote $z_i^{(m)}$ as the m-th imputation of a missing value in z_i. At site E:

1. Draw a random value U_i from a continuous uniform distribution $\mathcal{U}(0, 1)$.
2. Compute the weighted average of the F and $F + 1$ conditional predicted quantiles and assign:

$$z_i^{(m)} = (1 - \text{mod}) \cdot \hat{Q}_{z_i|\mathbf{w}_i}(F) + \text{mod} \cdot \hat{Q}_{z_i|\mathbf{w}_i}(F + 1) \tag{2}$$

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

In Stata: Impute the systematically missing variable



- **Set up MI environment**

```
mi set wide
mi register imputed z
```

- **Import coefficients and their variances**

```
mi_impute_from_get, b(siteC_b) v(siteC_v) ///
    colnames(y x a c d _cons) imodel(qreg)

mat ib = r(get_ib)
mat iV = r(get_iV)
```

- **Impute z multiple times**

```
mi impute from z , add(10) b(ib) v(iV) imodel(qreg)
```

```
External imputation using qreg              Imputations =       10
User method from                                  added =       10
Imputed: m=1 through m=10                       updated =        0

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
                 z |          0         6437      6437 |      6437
------------------------------------------------------------------
(Complete + Incomplete = Total; Imputed is the minimum across m
 of the number of filled-in observations.)
```

Overview

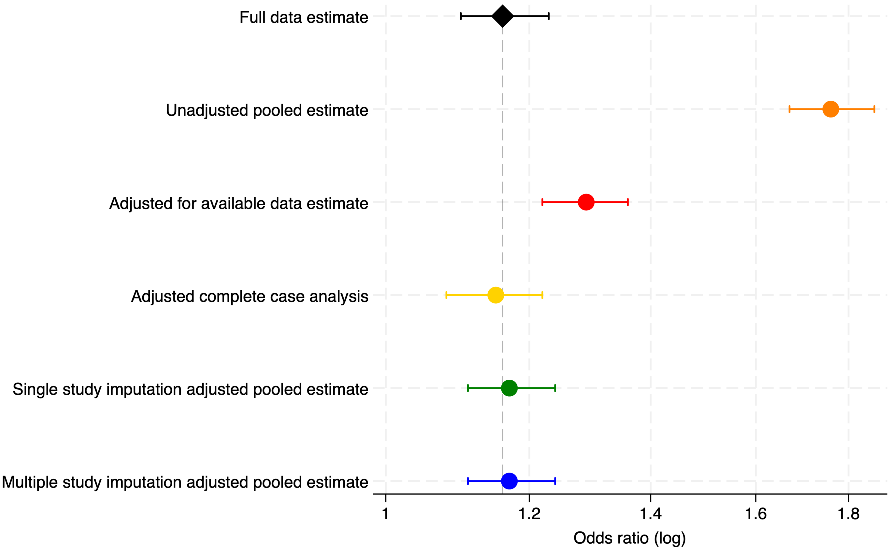
- Missing values in distributed data networks
- How does `mi impute from` work?
- Applied example
- Central assumptions and next steps

Maternal Antidepressants and Offspring neurodevelopmental disorders (NDD)

- 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's



Overview

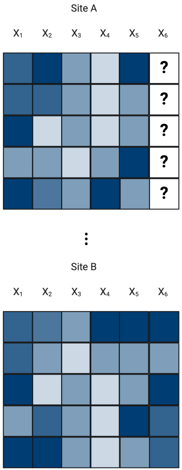
- Missing values in distributed data networks
- How does `mi impute from` work?
- Applied example
- Central assumptions and next steps

Cross-site imputation has two central assumptions

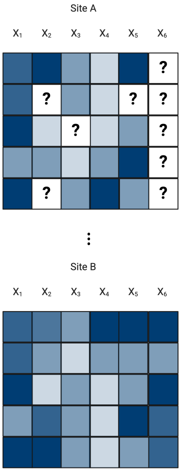
1. **Measurement assumption**: the variable we predict with observed data measures the same concept of the target variable we wish to impute (e.g., same measurement scale)
2. **Transportability assumption**: the association between the auxiliaries and the imputation target are transferable across sites. In other words, there is a ”common truth” to all sites, and each site represents a sample from that

Multivariate missing data

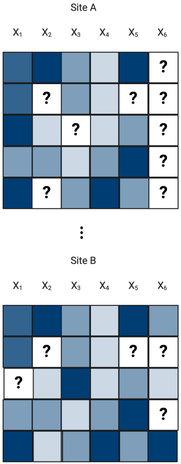
Univariate



Multivariate



Multivariate with incomplete auxiliaries



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 `mi impute from` command:

- can be used within the existing multiple imputation framework in Stata
- allows to import .txt and .xls files
- allows the use of logistic, multinomial logistic, and quantile regression for the imputation model
- has help documentation and a preprint

Future work may aim to:

- facilitate the use of more imputation commands
- integrate multivariate imputation

Acknowledgements

robert.thiesmeier@ki.se

`https://github.com/robertthiesmeier`

Acknowledgements

Nicola Orsini, Matteo Bottai, Scott Hofer (Karolinska Institutet, Sweden)
Viktor Ahlqvist (Aarhus University, Denmark)
Paul Madley-Dowd (University of Bristol, UK)
Sabina Murphy, Andrea Bellavia (Harvard University, USA)

Related work

1. 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

2. Thiesmeier R, Bottai M, Orsini N. Imputing missing data with external data. *Preprint*, 2024

3. Thiesmeier R, Madley-Dowd P, Orsini N, Ahlqvist V. Cross-site imputation for recovering variables without individual pooled data. *Preprint*, 2024