Handling missing data in meta-analysis of individual participant data with correlated mixed outcomes

Manuel Gomes

Centre for Statistical Methodology seminar

November 07, 2014

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Acknowledgments

✧ **UK MRC Early Career Felloswship in Economics of Health**

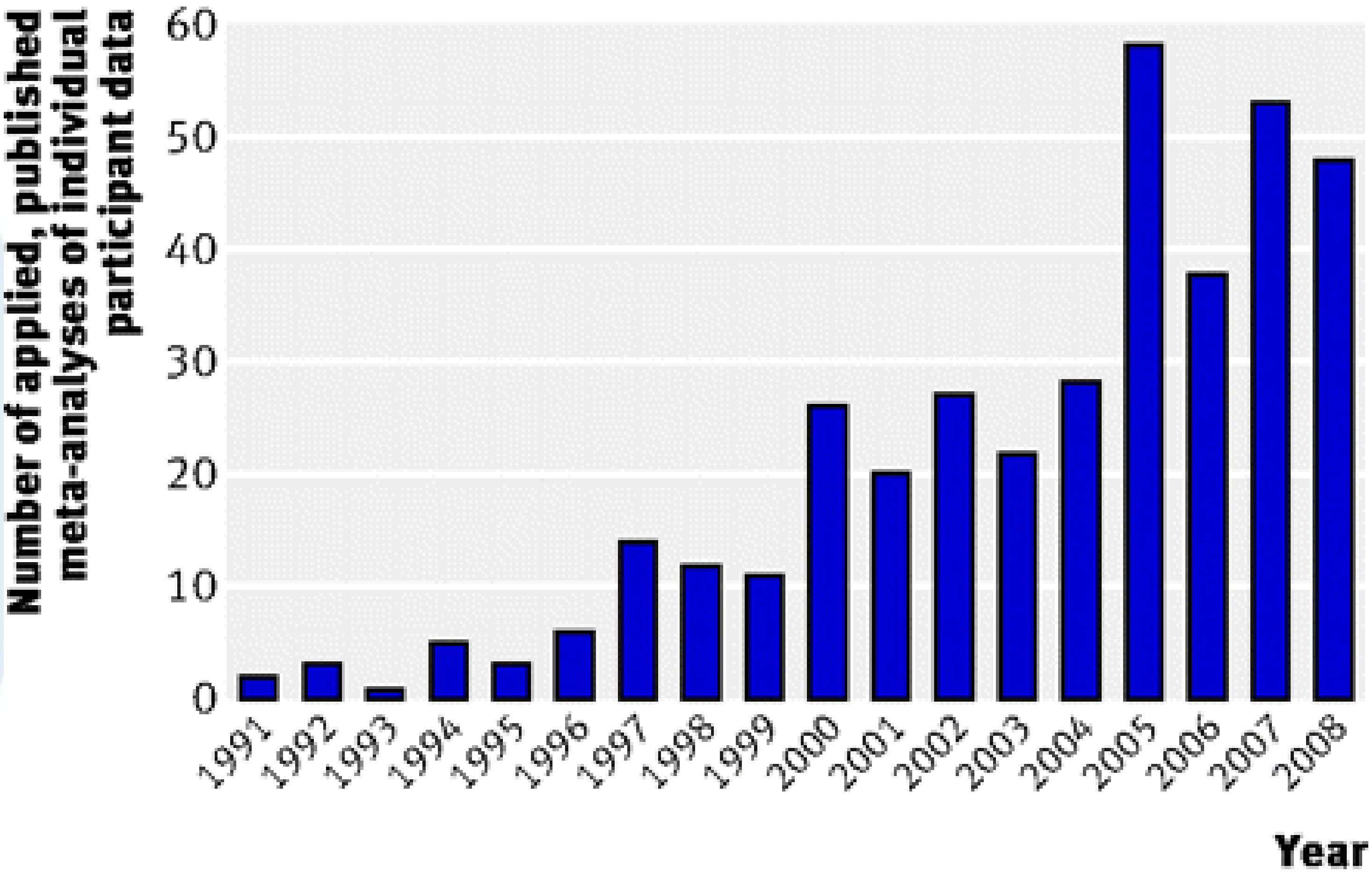
✧ **MDEpiNet Methodology Center at Harvard Medical School (**FDA/Chickasaw Nation Industries contract HHSF223201110172C; FDA grant 1U01FD004493-01**)**

- ✧ Laura Hatfield
- ✧ Sharon-Lise Normand

Overview

- Individual Participant Data (IPD) meta-analysis
- Methodological challenges
- A full-Bayesian approach
- Motivating example
- Some simulation results
- Discussion

IPD meta-analysis



Source: Riley et al 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 340 (7745):4521-525.

IPD meta-analysis

Key advantages of IPD meta-analysis:

- Consistent inclusion/exclusion criteria
- Analyses/modelling can be standardised across studies
- Potential for including additional confounders
- Missing data can be addressed using all available data

IN PRACTICE:

- Subgroup analysis
- Inference based on multiple outcomes

IPD meta analysis

Challenges:

- Outcomes are partially or completely unobserved for some studies
- Multiplicity of outcomes
- Between-study heterogeneity
- Small number of studies

A multivariate Bayesian hierarchical model for handling the missing data in IPD meta-analysis

A Bayesian approach

Full-Bayesian analysis

- Unknown parameters and missing values are estimated simultaneously, given the observed data.

$$f\left(\theta, Y^{mis} | Y^{obs}\right) = f\left(Y^{mis} \middle| Y^{obs}\right) f\left(\theta \middle| Y^{mis}, Y^{obs}\right)$$

- Flexible framework for addressing between-study heterogeneity, and modelling multiple mixed outcomes

- Offers additional advantages in the context of evidence synthesis when prior evidence is available

- Requires that all variables explaining missingness are included in the analysis model

A Bayesian approach

Bayesian hierarchical mixed model (Goldstein et al 2009)

Y_{ij}^{k} - k*th* continuous outcome, $k = 1, \ldots, K$; Y_{ij}^{l} - l*th* binary outcome, $l = 1, \ldots, L$

$$Y_{ij}^{k} = \mu_{ij}^{k} + \varepsilon_{ij}^{k}$$

$$Z_{ij}^{l} = \mu_{ij}^{l} + \varepsilon_{ij}^{l} \qquad P(Y_{ij}^{l} = 1) = P(z_{ij}^{l} > 0)$$

$$\mu_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 X_{ij} + u_j$$

$\varepsilon_{ij} \sim N(\mathbf{0}, \boldsymbol{\Omega}_{\varepsilon})$ where Ω_{ε} is the $k \times l$ level-1 covariance matrix with σ_l^2 constrained to 1

$u_j \sim N(\mathbf{0}, \boldsymbol{\Omega}_{u})$ where Ω_u is the $k \times l$ level-2 covariance matrix

A Bayesian approach

Diffuse priors (parameters constant across studies)

- **Level-2 covariance matrix**

$$\tau_k \sim N(0, 0.001)I(0,)$$
$$\tau_l \sim N(0, 0.001)I(0,)$$
$$\phi \sim Unif(-1, 1)$$

Inverse-Wishart prior could be used but uses a single parameter to control the precision of all elements of the matrix -> informative

- **Level-1 covariance matrix**

$$\sigma_k \sim N(0, 0.001)I(0,)$$
$$\rho \sim Unif(-1, 1)$$

- **Regression Coefficients**

$$\beta \sim N(0, 1.0E - 6)$$

Alternative methods

Alternative methods

Complete-case analysis (CCA)

- Patients with missing observations are dropped
- Assumes data are MCAR
- Typically leads to biased/imprecise results

Alternative methods

Multiple Imputation (MI)

- Each missing value is replaced by a set of plausible values drawn from the posterior distribution of missing data given the observed
  - Imputation model is estimated separately from the analysis model (allows for the inclusion of ‘auxiliary variables’)
  - Recognises the uncertainty associated with the missing data and the estimation of the imputed values
  - Implementation is relative simple and available in a wide range of software

Alternative methods

Multiple Imputation via fully-conditional specification

- Each incomplete variable is imputed iteratively.

Let $y_1^{(t-1)}$ and $y_2^{(t-1)}$ be the initial values. For each iteration t:

$$\theta_1^{(t)} \sim p(\theta_1)\, p(y_1^{obs} \,|\, y_2^{(t-1)}, X, \theta_1) \qquad y_1^{mis(t)} \sim p(y_1^{mis} \,|\, y_2^{(t-1)}, X, \theta_1^{(t)})$$

$$\theta_2^{(t)} \sim p(\theta_2)\, p(y_2^{obs} \,|\, y_1^{(t-1)}, X, \theta_2) \qquad y_2^{mis(t)} \sim p(y_2^{mis} \,|\, y_1^{(t-1)}, X, \theta_2^{(t)})$$

Typically after 10 to 20 iterations, y_1^{mis} and y_2^{mis} are imputed from the posterior distribution as follows:

y_1^{mis} is drawn from $p(y_1^{mis} \,|\, y_2^*,\ X, \theta_1^*)$

y_2^{mis} is drawn from $p(y_2^{mis} \,|\, y_1^*,\ X, \theta_2^*)$

Methodological intrigue

- In standard (non-hierarchical) settings, MI via FCS approximates the posterior distribution implied by the joint model (Gelman et al 2012)

- Previous work comparing JM with MI on correlated mixed outcomes found similar performance (He and Belin 2014)

- Unclear whether this holds in hierarchical settings:

  - Correlation structure is not explicitly partitioned between patient and study-level components

  - Marginal distribution may provide information about the conditional distribution. E.g., General location model

Motivating example

Case study

Analysis of five randomized controlled trials (N=5273):

- **Aim:** To compare cardiac resynchronization therapy (CRT) versus CRT combined with implantable cardioverter defibrillator (CRT-D) for treating chronic heart failure

- **Outcomes:**
  - Mortality
  - Functional: NYHA Class and 6-minute walk
  - Quality-of-life: Minnesota Living with Heart Failure questionnaire

- **Key research question**:

  Is the treatment effect different for males versus females?

Case study

Data completeness: complete cases: 2725 (52%)

Outcome	Mortality (5% missing)	NYHA class (15% missing)	6-min walk (22% missing)	Quality of Life (44% missing)
Study 1 (N=490)	✓✗	✓✗	✓✗	✓✗
Study 2 (N=555)	✓	✓✗	✓✗	✓✗
Study 3 (N=1798)	✓	✓✗	✓✗	✗
Study 4 (N=610)	✓	✓✗	✓✗	✓✗
Study 5 (N=1820)	✓✗	✓✗	✓✗	✗

✓: fully-observed;
✓✗: partially missing
✗: completely missing

Case-study: analysis model

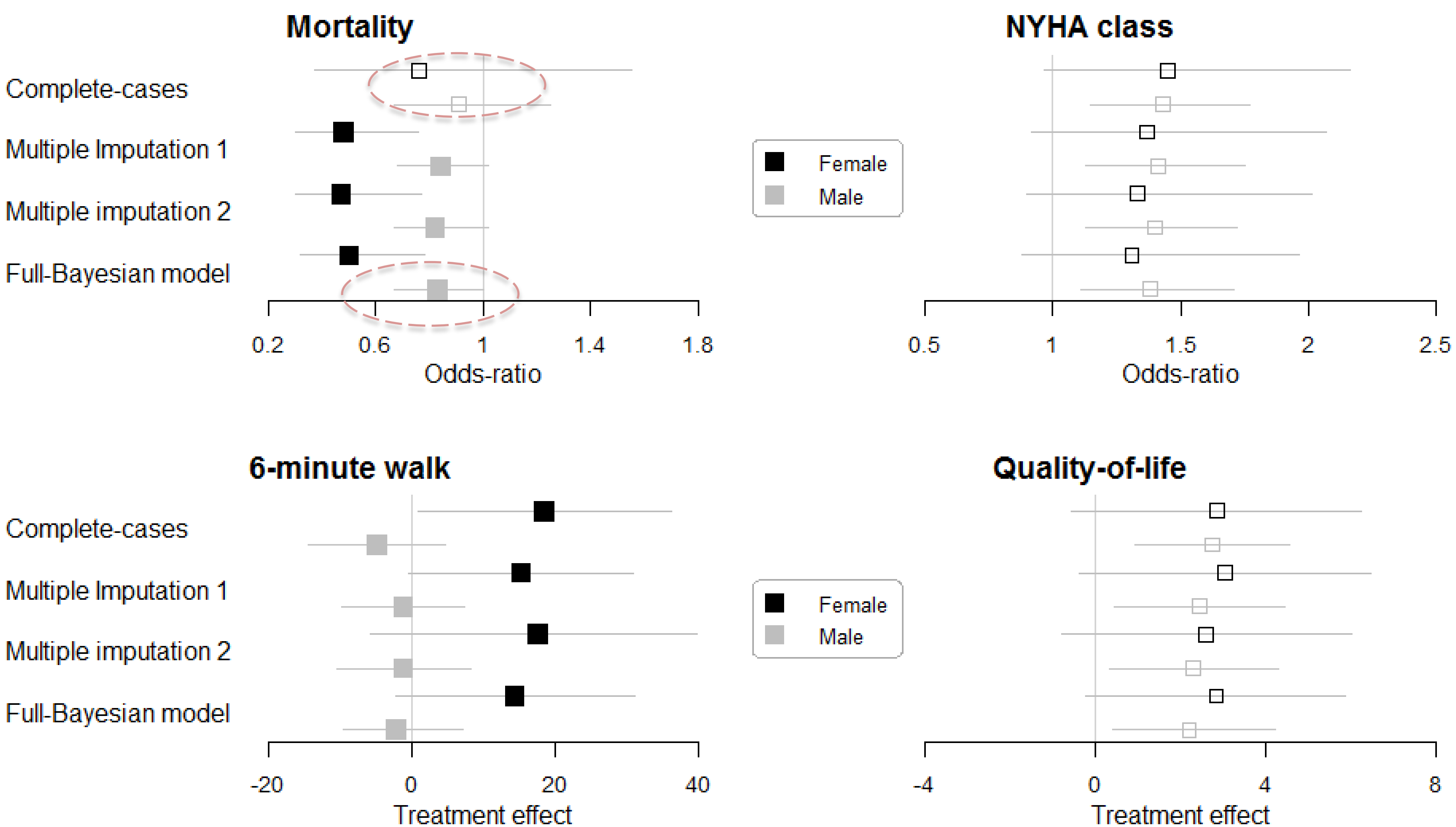
Multilevel mixed model (2 binary, 2 continuous)

$$
\begin{cases}
Z_{ij}^{death} \sim N(\mu_{ij}^{1}, 1) & P\left(death_{ij} = 1\right) = \mathrm{P}(z_{ij}^{death} > 0) \\
Z_{ij}^{nyha} \sim N(\mu_{ij}^{2}, 1) & P\left(nyha_{ij} = 1\right) = \mathrm{P}(z_{ij}^{nyha} > 0) \\
walk_{ij} \sim N(\mu_{ij}^{3}, \sigma_3^2) & \\
qol_{ij} \sim N(\mu_{ij}^{4}, \sigma_4^2) &
\end{cases}
$$

$$\mu_{ij}^{k} = \beta_0^k + \beta_1^k treat_{ij} + \beta_2^k\, sex_{ij} + \beta_3^k\, treat_{ij} * sex_{ij} + u_j^k$$

$$u_j^k \sim N(0, \Omega_u) \qquad k = 1, \dots, 4$$

Case-study: results



Case study: results

<u>Summary:</u>

1. Unprincipled methods for handling missing data

– <u>Complete-case analysis</u> with multiple outcomes is inefficient and leads to different inferences on mortality

– <u>Available-case analysis</u> could be used, BUT
  - Inconsistent sample across outcomes
  - No correlation between outcomes accounted for
  - Still assumes MCAR

2. Principled methods for handling the missing data

- Led to same inferences about the differential treatment effect
- CIs somewhat narrower for joint Bayesian model

Some simulation results

Simulations: data generation

Covariates

$$X_{ij}^{k} \sim N(0,1) \qquad \text{k} = 1, \dots, 4$$

Outcomes (multivariate distribution via copulas)

$$Y_{ij}^{1} \sim N(\mu_{ij}, \sigma_1^2) \qquad Y_{ij}^{2} \sim Bern(\pi_{ij})$$

$$\mu_{ij} = 1 + 1T_{ij} + 0.5X_{1,ij} + 0.5X_{2,ij} + u_j^1$$

$$logit(\pi_{ij}) = -0.5 - 0.1T_{ij} + 0.2X_{3,ij} + 0.2X_{4,ij} + u_j^2$$

$$\begin{pmatrix} u_j^1 \\ u_j^2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ & 1 \end{pmatrix}\right)$$

Simulations: implementation

1. Scenarios differed according to:

- **No. of studies:** 5 and 20
- **Correlation between outcomes:** 0.2 and 0.7
- **% Missing data:** 20% and 50%
- **Missingness data mechanism**
    - sporadically missing (MAR on X, MAR on X and Y)
    - systematically missing (MCAR)

2. Implementation

- **1000 replications**
- **Parameter of interest**: treatment effect on binary & continuous outcomes
- **Performance metrics:** bias, mean square error and joint CI coverage

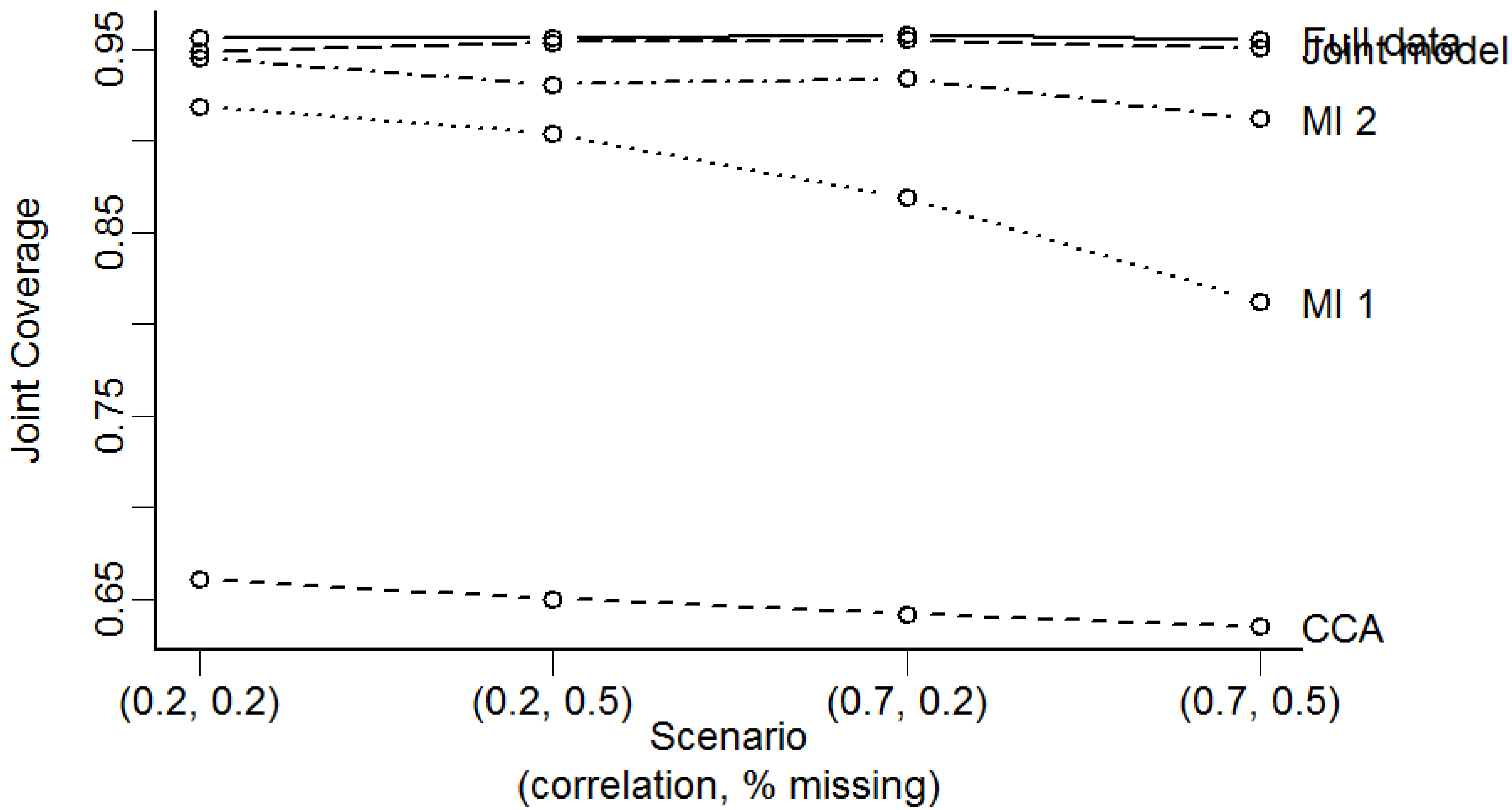
Simulations: results 1

Scenario: 20 studies, corr between outcomes=0.7, % missing=0.5, MAR on X

Method	Bias (%)		Root MSE		Joint coverage
	beta.Y1	beta.Y2	beta.Y1	beta.Y2	
Full data	0.0	1.2	0.040	0.075	0.958
Complete cases	10.1	34.9	0.121	0.128	0.503
MI 1	0.1	7.4	0.054	0.103	0.944
MI 2	0.1	6.1	0.049	0.100	0.947
Joint Model	0.0	3.9	0.040	0.083	0.957

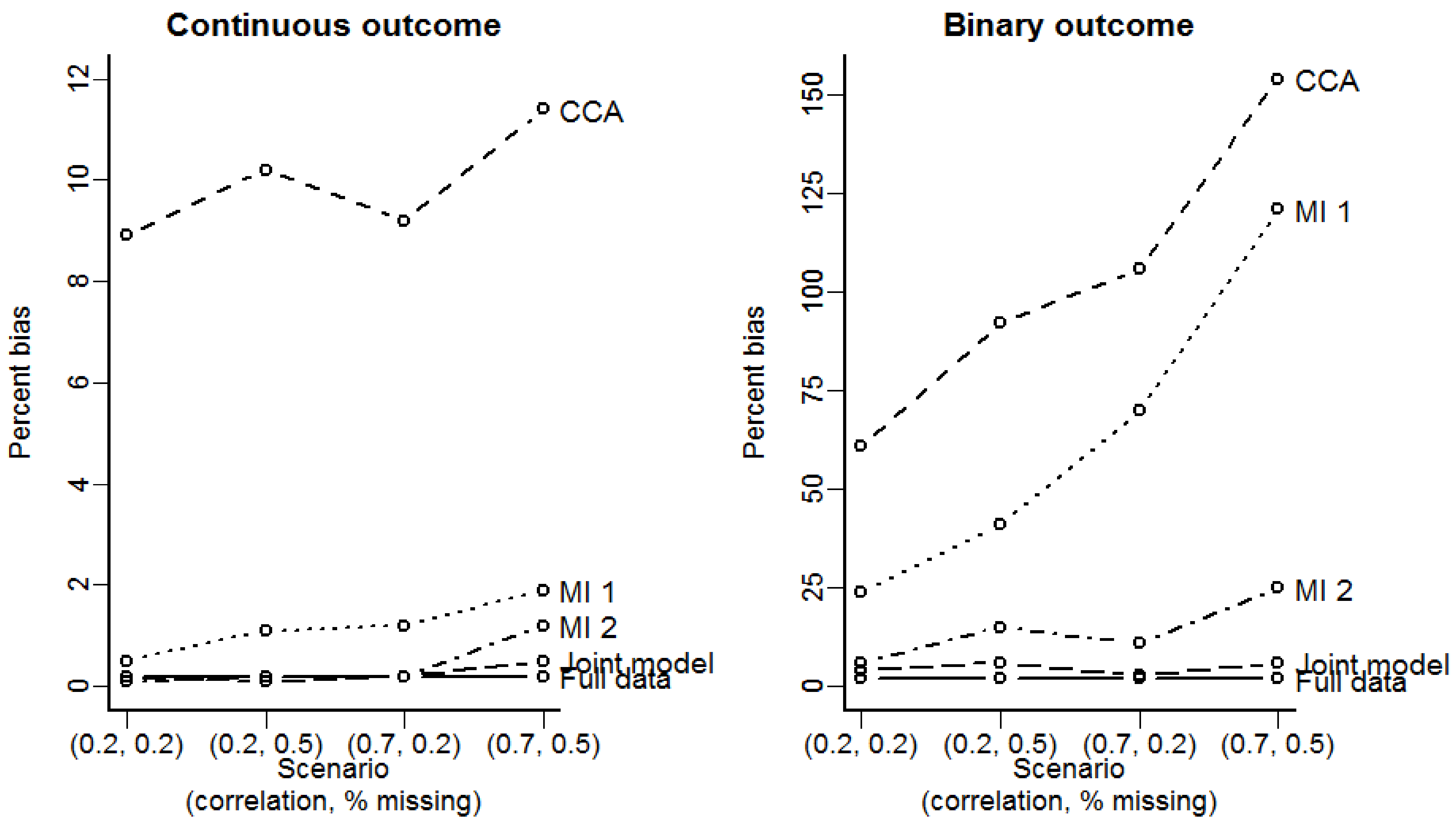
Simulations: results 2

Scenario: 5 studies, corr=0.7, % missing=0.5, MAR on X and Y



Simulations: results 2

Scenario: 5 studies, corr=0.7, % missing=0.5, MAR on X and Y



Discussion 1

- Similar results for scenarios with systematically missing data
- Bayesian approach performed well across all scenarios
- MI approach provided a decent alternative for most scenarios

Discussion 2

- Improvements to current implementation of MI via FCS
- Joint imputation model
- MI and two-stage meta-analysis with very few studies

Limitations and further work

- No missing covariates
- Fairly simple covariance structures

Ongoing work

- Assess the relative merits of the Bayesian approach under MNAR