Practical advantages of using R for handling missing data in HTA studies

R for HTA annual workshop

1-2 July 2021

Manuel Gomes

University College London

Outline

I will cover:

- Missing data issues in HTA (CEA) studies using IPD
- Advantages of using R in three HTA settings:
  - Hierarchical studies
  - Joint modelling
  - Missing not at random outcomes

This talk will not include:

- Case studies primarily based on modelling or aggregate data.
- Summary of all relevant R packages to HTA users facing missing data problems

Setting 1 - Hierarchical studies

- Hierarchical structure must be accounted for in the missing data model (as is in the substantive model)

- Probability of observing the data is likely to be more similar within groups/clusters (e.g. GP practices)
  - Patient characteristics more similar within those groups
  - Data collection efforts may differ across sites (clusters)

- Missing data methods that ignore clustering will lead to:
  - Imprecise cost-effectiveness estimates
  - Biased results if cluster size is informative - cost accumulation or treatment effectiveness changes with no. patients recruited to cluster (Gomes et al 2013)

Multilevel MI

Non-hierarchical MI model:

$$\begin{aligned} c_{ij}{}^{miss} &= \boldsymbol{\beta}^c X_{ij} + \boldsymbol{\gamma}^c Z_j + \varepsilon_{ij}^c \\ e_{ij}{}^{miss} &= \boldsymbol{\beta}^e X_{ij} + \boldsymbol{\gamma}^e Z_j + \varepsilon_{ij}^e \end{aligned} \qquad \begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix} \right)$$

- Between-cluster variation explained by **Z** can be account for

Multilevel MI

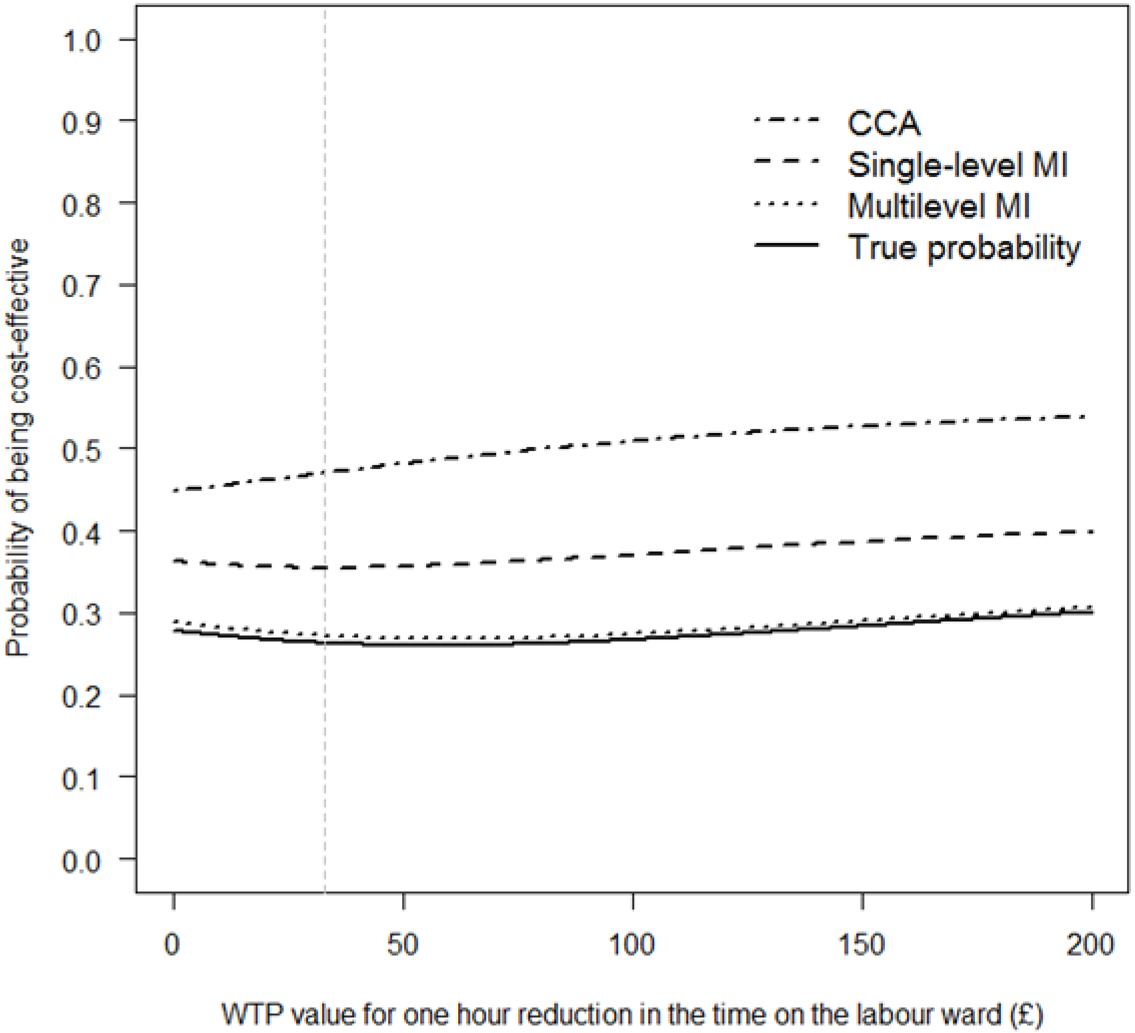
$$\begin{aligned} c_{ij}{}^{miss} &= \boldsymbol{\beta}^c X_{ij} + \boldsymbol{\gamma}^c Z_j + u_j^c + \varepsilon_{ij}^c \\ e_{ij}{}^{miss} &= \boldsymbol{\beta}^e X_{ij} + \boldsymbol{\gamma}^e Z_j + u_j^e + \varepsilon_{ij}^e \end{aligned} \qquad \begin{aligned} \begin{pmatrix} \varepsilon_{ij}^c \\ \varepsilon_{ij}^e \end{pmatrix} &\sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_e \\ & \sigma_e^2 \end{pmatrix} \right) \\ \begin{pmatrix} u_j^c \\ u_j^e \end{pmatrix} &\sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \phi\tau_c\tau_e \\ & \tau_e^2 \end{pmatrix} \right) \end{aligned}$$

- Bayesian hierarchical models would also be suitable (Diaz-Ordaz et al 2014)

Maternity study - Gomes et al 2013

- CEA of intervention to improve diagnosis of active labour in women having 1st child.

- Cluster trial
  - Few clusters (14 maternity units)
  - High within-cluster correlation (e.g ICC=0.14)
  - Re-analysed the data by simulating different missing data scenarios



Why R?

- Implementation of Multilevel MI outside R is challenging
  - Requires specialist software - e.g. REALCOM-Impute macros for MLwiN
  - Stata ‘mi impute’ currently does not allow for clustering
  - One can call REALCOM-Impute from Stata but prone to issues

- Several packages to implement multilevel MI
  - `Pan` - implements multilevel MI based on multivariate mixed model (Schafer and Yucel 2002)
  - `Mice` – can include random-effects, but less clear how the full hierarchical structure is handled when imputing non-Gaussian outcomes
  - `Jomo` – more recent package to handle joint hierarchical MI models

- Flexible platform to run Bayesian hierarchical models (More on this in Andrea’s talk)

Example R code

- ### Using mice package ##
- data0<-subset(data, arm==0, select=c(qaly,total_cost,cluster,agecat,eco_status,english,sizecl,bqaly,epds_6we,epds_6mo))

- ini <- mice(data0, maxit=0) #initial values
- pred <- ini$pred
- # Select variables to imputation model
- pred[1,] <- c(0, 1, -2, 1, 1, 1, 1, 1, 1, 1) # 1=predictor; -2=cluster; 0=variable to be imputed
- pred[2,] <- c(1, 0, -2, 1, 1, 1, 1, 1, 1, 1)

- imp0 <- mice(data0, m=M, meth=c("2l.pan","2l.pan", rep("",8)), seed=1710, pred=pred, maxit=5)

- Other options: 2l.norm, 2l.bin, 2l.jomo, 2lonly.norm (…)

Setting 2 - Joint modelling

- Joint modelling is central to CEA
  - Typically CEAs are required that costs and outcomes are jointly modelled
- Other settings also require joint modelling
- Individual patient data meta-analysis
  - Receiving increasing attention in HTA
  - Consistent inclusion/exclusion criteria
  - Analysis can be standardised across studies
  - Consider information beyond that included in original publication
  - **More plausible assumptions about the missing data**

Case study

FDA-commissioned IPD meta-analysis of cardiac devices

Aim: synthesise evidence from 5 RCTs (N=5273) on cardiac resynchronisation (CRT) alone versus CRT combined with cardio defibrillator for chronic heart failure

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 ;

IPD random-effects meta-analysis

Joint hierarchical model (2 binary, 2 continuous)

$$\begin{cases} Z_{ij}^{death} = \mu_{ij}^{1} + \varepsilon_{ij}^{1} \\ Z_{ij}^{nyha} = \mu_{ij}^{2} + \varepsilon_{ij}^{2} \\ walk_{ij} = \mu_{ij}^{3} + \varepsilon_{ij}^{3} \\ qol_{ij} = \mu_{ij}^{4} + \varepsilon_{ij}^{4} \end{cases}$$

$$P\left(death_{ij} = 1\right) = \mathrm{P}(Z_{ij}^{death} > 0)$$

$$P\left(nyha_{ij} = 1\right) = \mathrm{P}(Z_{ij}^{nyha} > 0)$$

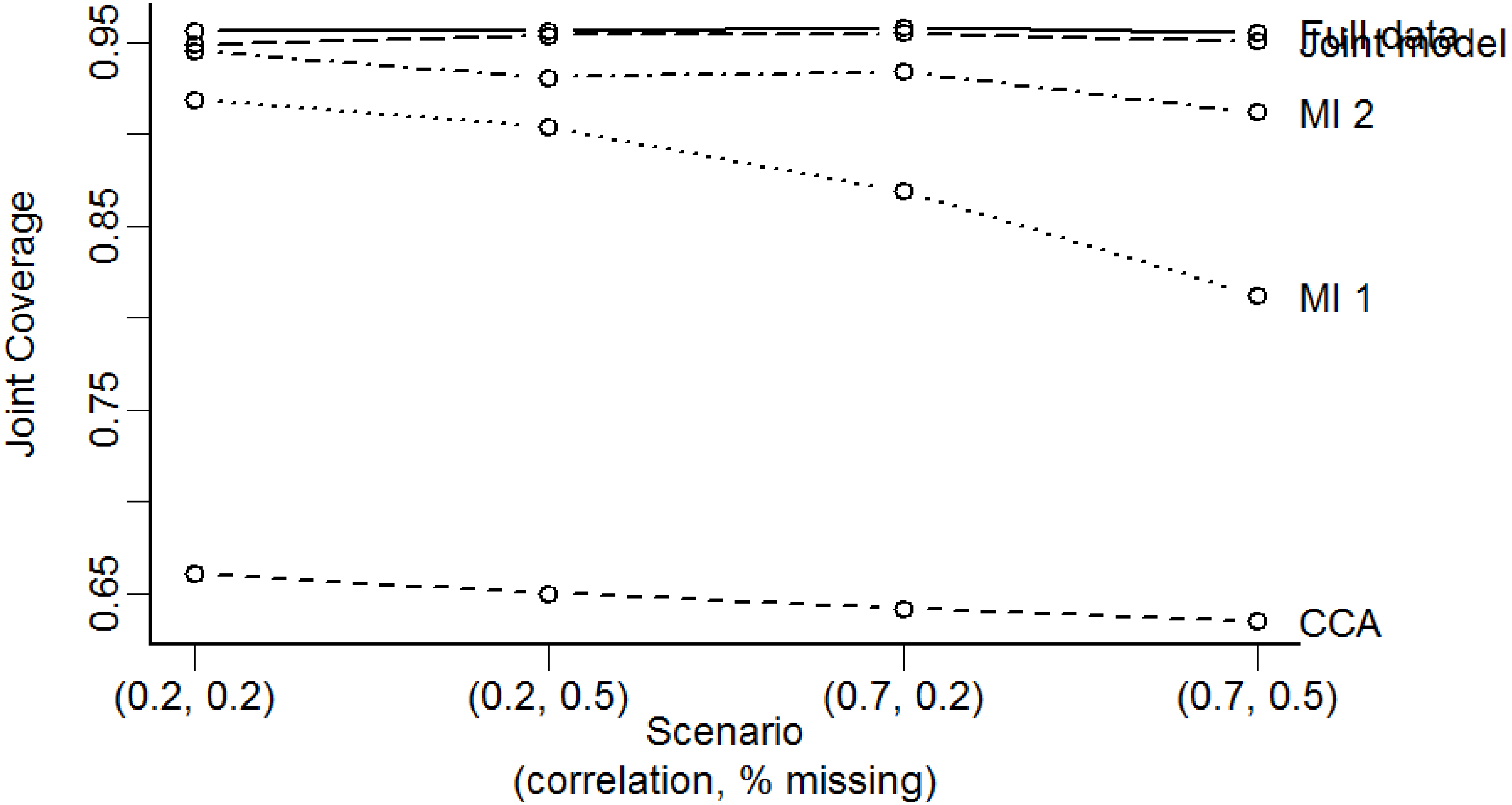
$$\mu_{ij}^{k} = \beta_0^k + \beta_1^k treat_{ij} + \beta_2^k\, sex_{ij} + \beta_3^k\, treat_{ij} * sex_{ij} + u_j^k$$

$\varepsilon_{ij}^{k} \sim N(0, \Omega_\varepsilon)$ $u_j^k \sim N(0, \Omega_u)$ $k = 1, \dots, 4$

RA study results

Compared joint versus chained equations MI (Gomes et al 2016)

- Correlation between outcomes at study-level not properly accounted for by the chained equations approach



Why R?

- Option to implement multivariate normal (MVN) MI was not available in Stata (back in 2013)
  - Now we can use `mi impute mvn` option
  - Clustering not allowed for
  - Again we'd have to use REALCOM-Impute macro (either in MLwiN or Stata)
- More sophisticated packages to conduct multilevel MI
  - For example, `jomo` package allows distinct imputation models for missing variables at patient versus study level
- Further flexibility to undertake IPD meta-analysis
  - Bringing data together (from different studies) is straightforward
  - Bayesian methods for evidence synthesis

Setting 3 - Missing not at random

- In many CEA settings, the chances of observing the data tend to be associated with the underlying **unobserved values**

- **For example**, patient-reported outcomes are widely used for assessing the benefits of health interventions (e.g. NICE, WHO), but are prone to missing data and unlikely to be MAR

- The chances of patients completing health questionnaires are typically related to their true health status, i.e. data are **missing not at random (MNAR)**

Selection models

- **Selection models** usually involve estimating the missing data and analysis models jointly

$$Y_i = \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$logit(P(R_i = 1)) = \gamma Z_i + \boldsymbol{\alpha} Y_i \qquad R_i = 1 \text{ if } Y_i \text{ is observed, } 0 \text{ otherwise}$$

Where the missing data model is a function of MNAR outcome.

This can be estimated in many ways (examples in CEA/econometrics)

- Heckman 2-step approach (Heckman 1976)
- MI (Gomes et al 2020)
- Copula models (Gomes et al 2019)
- Bayesian analysis (Mason et al 2021)

Pattern mixture models

- **Pattern mixture models** address MNAR by allowing for differences between the distribution of observed and unobserved data

$Y_i \sim N(\mu_i + \delta(1 - R_i),\ \sigma^2)$ $R_i = 1 \text{ if } Y_i \text{ is observed, } 0 \text{ otherwise}$

- Where the distribution of unobserved values differs from that of observed values by δ

This can be estimated in many ways (examples in CEA)

- Bayesian analysis (Mason et al 2018)
- MI (Leurent et al 2018)

Why R?

- Natural framework to conduct Bayesian analysis
  - E.g. using JAGS or Stan
  - Either selection or pattern mixture approaches
  - Flexible to handle non-Normal (and correlated) cost-effectiveness endpoints
  - Mason et al 2018 and 2021 provide R code for handling MNAR

- Flexibility offered for copula selection models (e.g. not available in Stata or SAS)
  - Wide range of non-Gaussian outcome distributions
  - Different copula functions (to reflect the dependence between non-response and the outcome)
  - `GJRM` package – R code provided in Gomes et al 2019

References

Schafer JL, Yucel RM (2002). Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics. 11:437-457

Gomes M, Diaz-Ordaz K, Grieve R, Kenward M (2013). Multiple imputation methods for cost-effectiveness analyses that use hierarchical studies: an application to cluster randomised trials. *Medical Decision Making*, 33 (8): 1051-1063.

Diaz-Ordaz K, Kenward MG, Grieve R. Handling missing values in cost-effectiveness analyses that use data from cluster randomised trials. J R Stat Soc Series A, 177 (2): 457-474.

Gomes M, Hatfield L, Normand SL (2016). Handling incomplete correlated continuous and binary outcomes in meta-analysis of individual participant data. *Statistics in Medicine,* 35 (21): 3676-89.

Heckman J. Sample bias as a specification error. *Econometrica*. 1979;47:153-162.

References

Leurent B, Gomes M, Faria R, et al. (2018) Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. *Pharmacoeconomics* 36 (8): 889-901

Mason A, Gomes M, Grieve R, Carpenter J (2018). A Bayesian framework for health economic evaluation in studies with missing data. *Health Economics*: 27 (11): 1670-1683.

Gomes M, Radice R, Camarena-Brenes J, Marra G (2019). Copula selection models for non-Gaussian outcomes missing not at random. *Statistics in Medicine,* 38 (3): 480-496.

Gomes M, Kenward MG, Grieve R, Carpenter JR (2020). Estimating treatment effects under non-ignorable missing data. *Statistics in Medicine* 39 (11): 1658-1674.

Mason A, Gomes M, Grieve R, Carpenter J (2021). Flexible Bayesian longitudinal models for cost-effectiveness analyses with informative missing data. *Health Economics* (in press).