MissingHE

An R package to handle missing data in trial-based health economic evaluations

Andrea Gabrio

Maastricht University (FHML)

Department of Methodology and Statistics

email: a.gabrio@maastrichtuniversity.nl

EuHEA Conference, Wien, 2024

Monday, July 1, 2024

Part I

Introduction to statistical modelling in HTA

Individual-level data in HTA

- Typically collected from clinical studies (e.g. RCTs) at multiple time points

		Demographics			HRQL data				Resource use data				Clinical outcome			
ID	**Trt**	Sex	Age	…	u_0	u_1	…	u_J	c_0	c_1	…	c_J	y_0	y_1	…	y_J
1	1	M	23	…	0.32	0.66	…	0.44	103	241	…	80	y_{10}	y_{11}	…	y_{1J}
2	1	M	21	…	0.12	0.16	…	0.38	1 204	1 808	…	877	y_{20}	y_{21}	…	y_{2J}
3	2	F	19	…	0.49	0.55	…	0.88	16	12	…	22	y_{30}	y_{31}	…	y_{3J}
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

$y_{ij} =$ Survival time, event indicator (eg CVD), number of events, continuous measurement (eg blood pressure), …
$u_{ij} =$ Utility-based score to value health (eg EQ-5D, SF-36, Hospital Anxiety & Depression Scale, …)
$c_{ij} =$ Use of resources (drugs, hospital, GP appointments, …)

Individual-level data in HTA

- Typically collected from clinical studies (e.g. RCTs) at multiple time points

		Demographics			**HRQL data**				**Resource use data**				**Clinical outcome**			
ID	**Trt**	Sex	Age	…	u_0	u_1	…	u_J	c_0	c_1	…	c_J	y_0	y_1	…	y_J
1	1	M	23	…	0.32	0.66	…	0.44	103	241	…	80	y_{10}	y_{11}	…	y_{1J}
2	1	M	21	…	0.12	0.16	…	0.38	1 204	1 808	…	877	y_{20}	y_{21}	…	y_{2J}
3	2	F	19	…	0.49	0.55	…	0.88	16	12	…	22	y_{30}	y_{31}	…	y_{3J}
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

y_{ij} = Survival time, event indicator (eg CVD), number of events, continuous measurement (eg blood pressure), …
u_{ij} = Utility-based score to value health (eg EQ-5D, SF-36, Hospital Anxiety & Depression Scale, …)
c_{ij} = Use of resources (drugs, hospital, GP appointments, …)

- Outcome measures evaluated over time, e.g. QALYs and total costs as

$$e_i = \sum_{j=1}^{J} \left(u_{ij} + u_{ij-1}\right) \frac{\delta_j}{2} \quad \text{and} \quad c_i = \sum_{j=1}^{J} c_{ij}, \ \left[\delta_j = \frac{\mathsf{T}_j - \mathsf{T}_{j-1}}{\text{Unit of T}}\right]$$

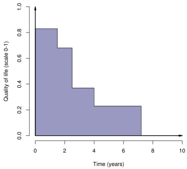
Individual-level data in HTA

- Typically collected from clinical studies (e.g. RCTs) at multiple time points

		Demographics			HRQL data				Resource use data				Clinical outcome			
ID	**Trt**	Sex	Age	...	u_0	u_1	...	u_J	c_0	c_1	...	c_J	y_0	y_1	...	y_J
1	1	M	23	...	0.32	0.66	...	0.44	103	241	...	80	y_{10}	y_{11}	...	y_{1J}
2	1	M	21	...	0.12	0.16	...	0.38	1 204	1 808	...	877	y_{20}	y_{21}	...	y_{2J}
3	2	F	19	...	0.49	0.55	...	0.88	16	12	...	22	y_{30}	y_{31}	...	y_{3J}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

y_{ij} = Survival time, event indicator (eg CVD), number of events, continuous measurement (eg blood pressure), ...
u_{ij} = Utility-based score to value health (eg EQ-5D, SF-36, Hospital Anxiety & Depression Scale, ...)
c_{ij} = Use of resources (drugs, hospital, GP appointments, ...)

- Outcome measures evaluated over time, e.g. QALYs and total costs as



QALY = "Area Under the Curve"

What should I be worried about?

- Potential **correlation** between costs & benefits
  - Strong positive/negative correlation — effective treatments are innovative and result from intensive and lengthy research
  - Methods: **Seemingly unrelated regression** (Zellner, 1962); **joint modelling**

What should I be worried about?

- Potential **correlation** between costs & benefits
  - Strong positive/negative correlation — effective treatments are innovative and result from intensive and lengthy research
  - Methods: **Seemingly unrelated regression** (Zellner, 1962); **joint modelling**

- **Non-normality** of the outcomes
  - Costs usually skewed and benefits may be bounded in [0;1] with possible spikes (e.g. zero costs)
  - Methods: **transformations**; **Generalised Linear Modelling**; **two-part/hurdle models**

What should I be worried about?

- Potential **correlation** between costs & benefits
  - Strong positive/negative correlation — effective treatments are innovative and result from intensive and lengthy research
  - Methods: **Seemingly unrelated regression** (Zellner, 1962); **joint modelling**

- **Non-normality** of the outcomes
  - Costs usually skewed and benefits may be bounded in [0;1] with possible spikes (e.g. zero costs)
  - Methods: **transformations**; **Generalised Linear Modelling**; **two-part/hurdle models**

- and of course **missing data**
  - May occur in multiple variables and substantially reduce the sample size
  - Methods: **No easy solution!**
  - Any method relies on **untestable assumptions**
  - Use a **principled approach** based on well-defined statistical model for the complete data, and explicit assumptions about missingness

Part II

Missing Data

Useful questions

- **How much** missingness?
  - If few variables and small rates (e.g.$< 5\%$) unlikely to affect results

Useful questions

- **How much** missingness?
  - If few variables and small rates (e.g.$< 5\%$) unlikely to affect results

- **Which** variables and patterns?
  - Outcomes vs predictors, dropout vs intermittent

Useful questions

- **How much** missingness?
  - If few variables and small rates (e.g.$< 5\%$) unlikely to affect results
- **Which** variables and patterns?
  - Outcomes vs predictors, dropout vs intermittent
- **Why** missingness occurred?
  - Random chance, individual characteristics observed/unobserved

Useful questions

- **How much** missingness?
  - If few variables and small rates (e.g.$< 5\%$) unlikely to affect results

- **Which** variables and patterns?
  - Outcomes vs predictors, dropout vs intermittent

- **Why** missingness occurred?
  - Random chance, individual characteristics observed/unobserved

- Different assumptions about the **mechanism** underlying missingness
- **Rubin's taxonomy** (Rubin, 1986) groups the mechanisms into:
  - Missing Completely At Random - does not depend on observed/unobserved data
  - Missing At Random - does not depend on unobserved data given the observed data
  - Missing Not At Random - depends on unobserved data given the observed data

Sensitivity analysis - MNAR

- Fully probabilistic approach (i.e. fundamentally Bayesian):
- Specify the joint distribution $p(y, m \mid \omega)$ using a **Pattern mixture model** approach:

$$p(y, m \mid \omega) = p\left(y \mid m, \omega^{\mathrm{PMM}}\right) p\left(m \mid \omega^{\mathrm{PMM}}\right)$$

- **Pattern mixture models**
  - A marginal model for the missingness patterns $p(m \mid \omega^{\mathrm{PMM}})$ and a conditional model for the response within each pattern $p(y \mid m, \omega^{\mathrm{PMM}})$
  - Intuitive to formulate assumptions for each pattern but difficult to fit with sparse data

Sensitivity analysis - MNAR

- Fully probabilistic approach (i.e. fundamentally Bayesian):
- Specify the joint distribution $p(y, m \mid \omega)$ using a **Selection model** approach:

$$p\left(m \mid y, \omega^{\mathrm{SM}}\right) p\left(y \mid \omega^{\mathrm{SM}}\right)$$

- **Selection models**
  - A marginal model for the response $p(y \mid \omega^{\mathrm{SM}})$ and the missing data mechanism $p(m \mid y, \omega^{\mathrm{SM}})$
  - Directly model the distribution of y but impact of missingness assumptions unclear

Missing data in HTA – Conclusions

- Often, little or no information is available about missingness

Missing data in HTA – Conclusions

- Often, little or no information is available about missingness
- Restrict the analysis to a single scenario (MAR) is unlikely to provide a realistic assessment of cost-effectiveness

Missing data in HTA – Conclusions

- Often, little or no information is available about missingness

- Restrict the analysis to a single scenario (MAR) is unlikely to provide a realistic assessment of cost-effectiveness

- **Selection** and **pattern mixture** models represent possible choices to perform sensitivity analysis to MNAR
  - Rely on **untestable assumptions** about the unobserved data
  - Useful to assess the robustness of the results to a range of **plausible** departures

Missing data in HTA – Conclusions

- Often, little or no information is available about missingness
- Restrict the analysis to a single scenario (MAR) is unlikely to provide a realistic assessment of cost-effectiveness
- **Selection** and **pattern mixture** models represent possible choices to perform sensitivity analysis to MNAR
  - Rely on **untestable assumptions** about the unobserved data
  - Useful to assess the robustness of the results to a range of **plausible** departures
- The Bayesian approach allows the incorporation of **external evidence** into the analysis for:
  - The selection of the assumptions to explore
  - The quantification of the impact of missingness on decision-making

Part III

missingHE: dealing with missing data in HTA

missingHE: package design

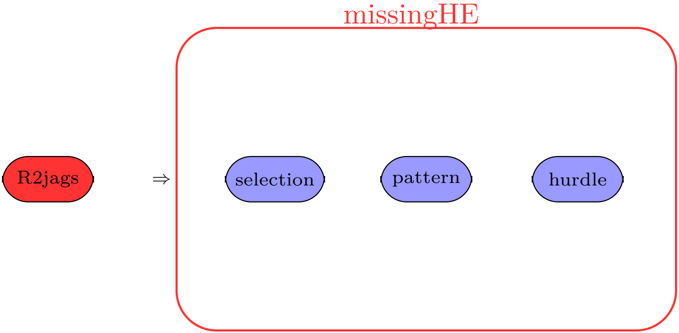
- The **missingHE** package provides different functions to fit Bayesian models for missing data in trial-based HTA

missingHE
R2jags
⇒

GitHub repository: https://github.com/AnGabrio/missingHE
CRAN repository: https://cran.r-project.org/web/packages/missingHE

missingHE: package design

- The **missingHE** package provides different functions to fit Bayesian models for missing data in trial-based HTA



GitHub repository: https://github.com/AnGabrio/missingHE
CRAN repository: https://cran.r-project.org/web/packages/missingHE

How to fit a model in missingHE

- **Example**: selection model
- The selection function takes several inputs
  - data: the dataframe object (containing e, c, t)
  - model.eff & model.cost: the formulae for the effect and cost model. Joint models are always formulated as $p(e, c) = p(e)p(c \mid e)$
  - model.me & model.mc: the formulae for the missing data mechanism for e and c
  - dist_e & dist_c: distributions assumed for e and c
  - type: the type of missing data mechanism (i.e. MAR/MNAR)
  - ...: optional inputs (e.g. MCMC iterations, user-defined priors, save model code, etc.)

```
> selection(data = MenSS, model.eff = e ~ u.0, model.cost = c ~ e,
+   model.me = me ~ age + e, model.mc = mc ~ age, type = ”MNAR”,
+   n.iter = 2000, dist_e = ”norm”, dist_c = ”gamma”, prior = my.prior)
```

How to fit a model in missingHE

- Start with assuming a joint model for the observed data $p(e, c \mid \theta)$ based on a Normal for e and a Gamma for c while also adjusting for u_0:

$$e_{it} \sim \text{Normal}\left(\phi_{iet}, \sigma^2_{ct}\right), \quad \phi_{iet} = \alpha_{0et} + \alpha_{1et} u_{0it}$$

$$c_{it} \sim \text{Gamma}\left(\frac{\phi^2_{ict}}{\sigma^2_{ct}}, \frac{\phi_{ict}}{\sigma^2_{ct}}\right), \quad \log(\phi_{ict}) = \alpha_{0ct} + \alpha_{1ct} e_{it}$$

- When specified, covariates are always included at the (conditional) mean level using appropriate link functions for both e and c models and must be fully-observed.

How to fit a model in missingHE

- **Example**: selection model
- Next, assume MNAR mechanism for e and MAR mechanism for c given age

$$m_{iet} \sim \text{Bernoulli}\left(\pi_{iet}\right), \quad \text{logit}(\pi_{iet}) = \gamma_{0et} + \gamma_{1et}\text{age}_{it} + \delta_e e_{it}$$

$$m_{ict} \sim \text{Bernoulli}\left(\pi_{ict}\right), \quad \text{logit}(\pi_{ict}) = \gamma_{0ct} + \gamma_{1ct}\text{age}_{it}$$

- When specified, covariates are always included at the (conditional) probability level using a logit link function for both m_e and m_c models and must be fully-observed.

How to fit a model in missingHE

- Informative priors on δ_e must be specified. By default **missingHE** uses a standard normal but hyperprior values can be changed using the optional argument
- Define a new list with object named delta.prior.e containing the new prior mean and sd values for δ_e (logit scale). For example:

```
> my.prior <- list(”delta.prior.e” = c(5, 1))
```

How to fit a model in missingHE

- Informative priors on δ_e must be specified. By default **missingHE** uses a standard normal but hyperprior values can be changed using the optional argument
- Define a new list with object named delta.prior.e containing the new prior mean and sd values for δ_e (logit scale). For example:

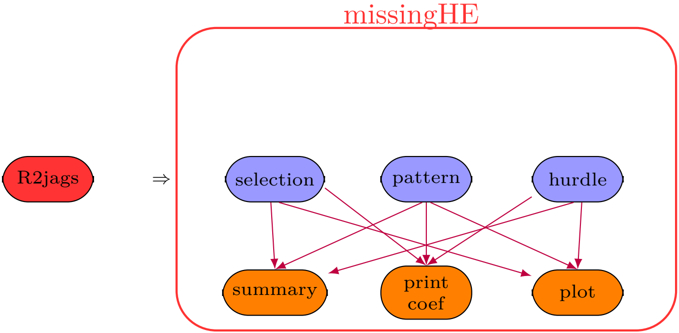
```
> my.prior <- list(”delta.prior.e” = c(5, 1))
```

- Pass my.prior as argument to the selection function and run

```
> NG.sel=selection(data = MenSS, model.eff = e ~ u.0, model.cost = c ~ e,
+   model.me = me ~ age + e, model.mc = mc ~ age, type = ”MNAR”,
+   n.iter = 2000, dist_e = ”norm”, dist_c = ”gamma”, prior = my.prior)
```

missingHE: package design

- The **missingHE** package provides different functions to fit Bayesian models for missing data in trial-based HTA



GitHub repository: https://github.com/AnGabrio/missingHE
CRAN repository: https://cran.r-project.org/web/packages/missingHE

missingHE: how to check posterior results

- We can check posterior summaries for μ_e and μ_c using the print or coef command

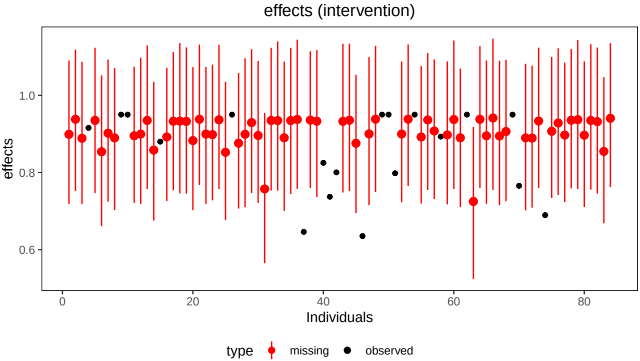
```
> print(NG.sel)
```

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
mu_e[1]	0.838	0.017	0.803	0.827	0.839	0.849	0.870	1.013	120
mu_e[2]	0.894	0.022	0.851	0.879	0.893	0.908	0.941	1.028	63
mu_c[1]	230.018	72.516	128.264	179.254	219.503	265.533	414.828	1.026	64
mu_c[2]	292.425	183.210	96.164	174.251	239.750	348.376	838.010	1.054	47

missingHE: how to check posterior results

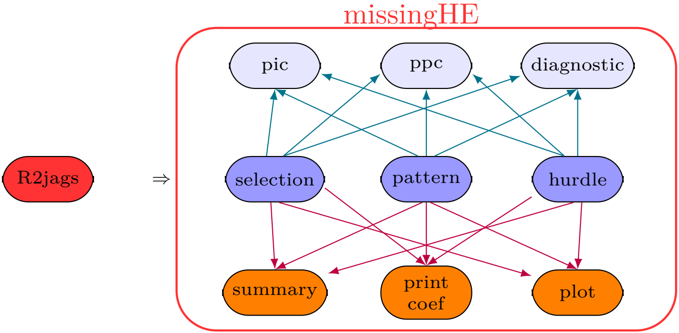
- We can check imputations by variable and arm using plot. For example, to display the imputed e in the reference $(t = 2)$ group we type

```
> plot(NG.sel, outcome = "effects_arm2")
```



missingHE: package design

- The **missingHE** package provides different functions to fit Bayesian models for missing data in trial-based HTA

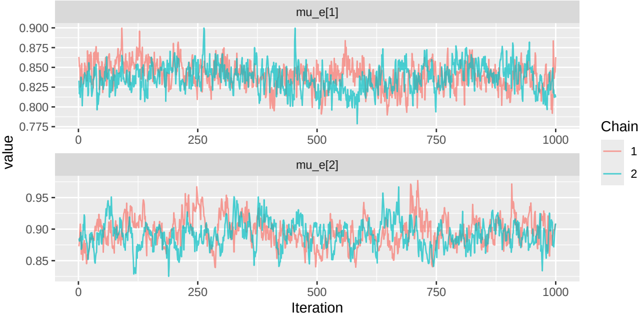


GitHub repository: https://github.com/AnGabrio/missingHE
CRAN repository: https://cran.r-project.org/web/packages/missingHE

missingHE: how to assess convergence and model fit

- Diagnostic tools for MCMC convergence (e.g. Rhat) can provide useful insights into potential issues of the algorithm
- Graphical MCMC diagnostics can be obtained for each model parameter. For example, we can use the diagnostic command to examine posterior traceplots for the mean effects by arm

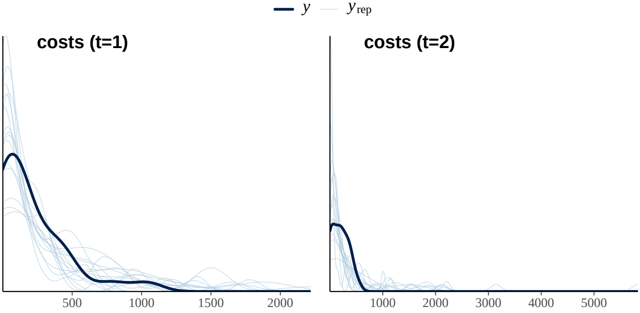
```
> diagnostic(NG.sel, type = "traceplot", param = "mu.e")
```



missingHE: how to assess convergence and model fit

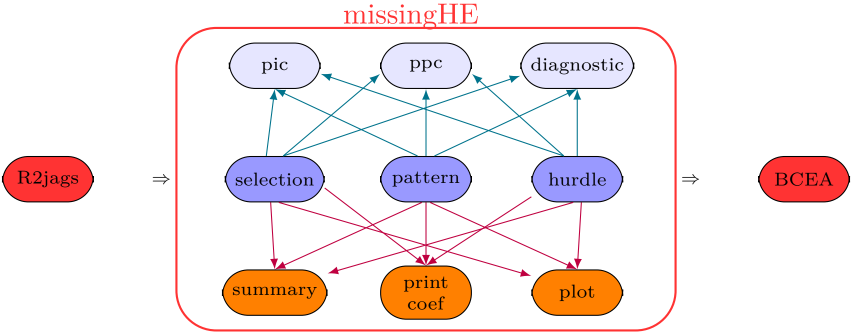
- Fit of the model assessed using Posterior Predictive Checks (PPCs)
- PPCs via statistics or graphical tools. For example, we can use the ppc command to examine histograms based on posterior densities (e.g. for costs).

```
> ppc(NG.sel, type = ”dens_overlay”, outcome = ”costs”, ndisplay = 15)
```



missingHE: package design

- The **missingHE** package provides different functions to fit Bayesian models for missing data in trial-based HTA

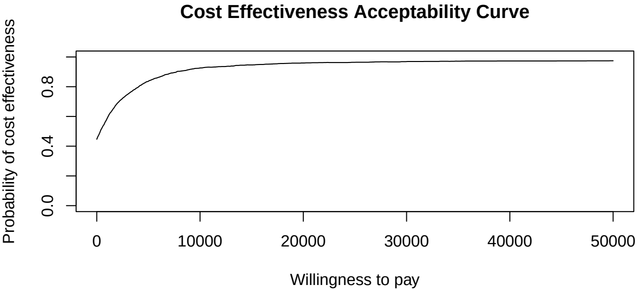


GitHub repository: https://github.com/AnGabrio/missingHE
CRAN repository: https://cran.r-project.org/web/packages/missingHE

missingHE: how to evaluate cost-effectiveness with BCEA

- Finally, standardised CEA output can be obtained by post-processing the results from the model fitted in **missingHE** using functions from the **BCEA** package (Baio, 2014)
- For example, CE acceptability curves can be computed via the function ceac.plot to the object cea stored inside the model output

```
> library(BCEA)
> ceac.plot(NG.sel$cea)
```



Part IV

Discussion and conclusions

Discussion

- Individual-level HTA data are subject to some complexities (**including missingness!**) that are typically ignored by the "standard" approach
- A Bayesian approach allows to increase model complexity to jointly account for these complexities with relatively little expansion to the basic model
- MAR can be used as reference assumption but **plausible** MNAR departures should be explored in sensitivity analysis to assess robustness of results
- Possible to expand the framework to a longitudinal setting to handle missingness more efficiently (Gabrio et al. (2022). *RJSS: Series A*, 607-629)

missingHE: what to know and how to use it

- Specifies a set of pre-defined Bayesian models using the **R2jags** package
- Is linked to the **BCEA** package, which provides summary HTA results
- A comprehensive guide to the use of the package and the interpretation of the output is provided through a series of online **vignettes**
  - *Introduction*: a guide to the use of the main functions of the package
  - *Fitting MNAR models*: how to specify MNAR assumptions for each type of modelling approach available
  - *Model Customisation*: how to customise the model in different ways
- Instructions on how to use **missingHE** to fit and assess different types of models can be accessed by typing help on the different functions of the package
- A short course and code are available on my GitHub page