Clinical prediction models

Session 11
Dealing with missing data

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br
2018

Objectives

- Brief review the theoretical background on mechanisms of missingness of predictor values
- Comment how these missingness may affect the modelling process.
- Show examples on imputation methods as a solution
- This session is not intended to exhaust the missing/imputation topic

Problems

- Missing data are a common problem
- Standard statistical software for regression analysis deletes subjects with any missing data on any predictor before analysis
- Therefore, numbers of subjects may vary per analysis as different predictors are explored
- Complete case analysis are hence statistically inefficient

Rationale

- One must assume that true predictor values are hidden by the missing values.
- One must understand that imputations is not a “good guess” of the missing data, rather a good use of the available data.
- Evidence points to greater bias in predictions in complete case analysis when compared to analysis with imputed dataset.

Rationale

Table 7.1 Hypothetical missing data pattern: 250 subjects have partially complete data (missing data indicated with .), and 250 have fully complete data (indicated with *X*)

ID	*X*1	*X*2	*X*3	*X*4	*X*5	*Y*
1	.	X	X	X	X	X
…	.	X	X	X	X	X
50	.	X	X	X	X	X
51	X	.	X	X	X	X
…	X	.	X	X	X	X
100	X	.	X	X	X	X
101	X	X	.	X	X	X
…	X	X	.	X	X	X
150	X	X	.	X	X	X
151	X	X	X	.	X	X
…	X	X	X	.	X	X
200	X	X	X	.	X	X
201	X	X	X	X	.	X
…	X	X	X	X	.	X
250	X	X	X	X	.	X
251	X	X	X	X	X	X
…	X	X	X	X	X	X
…	X	X	X	X	X	X
500	X	X	X	X	X	X
Total	450	450	450	450	450	500

Steyerbeg. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer in 2009.

Rationale

- Basics
  - If each predictor has 10% missing data and that each patient has at most 1 missing value
  - Information available is 250 complete cases (250 × 5=1,250) + 250 incomplete cases (250 × 4=1,000) = 90% of the required data
  - Complete case data will use only 250/500 of patients in data
  - 10% missing -> 50% patients discarded

Rationale

- Basics
  - For example, one may wish to compare nested models, or adjust analysis and have an idea of the adjusted effect from univariable to multivariable
  - In two models conducted with missing data, it is then impossible to infer whether differences in odds ratios, p values or R2 arose because of true differences, because of correlation between the predictors or because of a selection of subjects due to missing values

Missing mechanisms

- Depending of the imputation strategy, the mechanism is not that relevant.
- In health data the mechanism is usually not at random.

Table 7.2 Three types of missing data mechanisms

Label	Missing mechanism	Description
MCAR	Missing completely at random	Administrative errors, accidents
MAR	Missing at random	Missingness related to known patient characteristics, time or place ("MAR on x"), or to the outcome ("MAR on y")
MNAR	Missing not at random	Missingness related to the value of the predictor, or to characteristics not available in the analysis

Steyerbeg. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer in 2009.

Examples of bias

- A correlation between missingness of a predictor and the outcome poses a serious problem in predictive modelling.
- If an association between missingness of predictors *X* and outcome *Y* is noted in a prospective study, the explanation must be through other predictors.
- MAR on *y* for only one predictor is sufficient to bias coefficients of all predictors.

Steyerbeg. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer in 2009.

Bias due to missing data

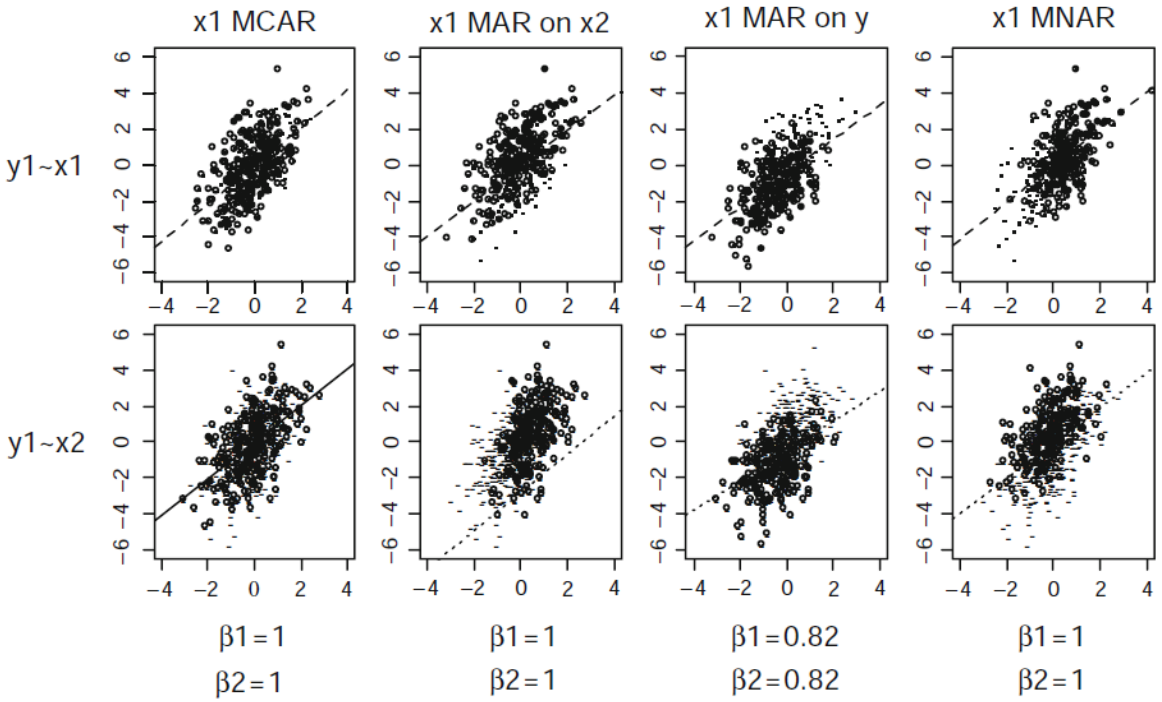

Fig. 7.1 Effect of missing values on estimated regression coefficients $\beta 1$ and $\beta 2$ in the model $y \sim X1+X2$. Original data are marked as "dot" and "dash" for $X1$ and $X2$, respectively. Complete data under MCAR, MAR, and MNAR are marked with a *circle*. Plots show results for $n = 500$; expected values for $\beta 1$ and $\beta 2$ are shown under the graphs (based on $n = 100{,}000$)

Steyerbeg. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer in 2009.

Imputation

- Imputation methods substitute the missing values by plausible values
- As the relation with the outcome is the main source of bias, always include the outcome in the imputation process
- Consider correlated predictors in the imputations process even if one of them is not going to be modeled: e.g. Hct <-> Hg

Imputation

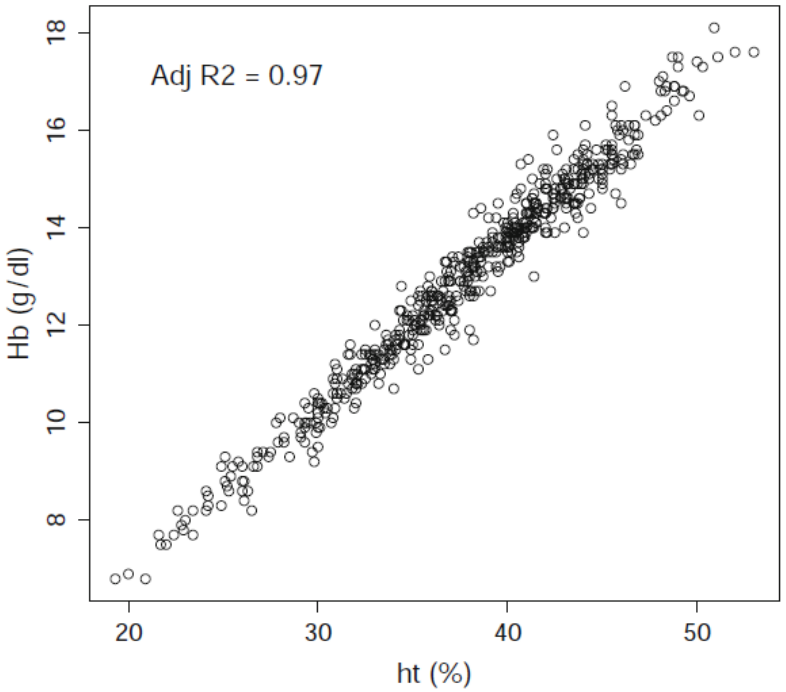

Fig. 7.3 Correlation between haematocrit (ht) and haemoglobin (Hb) in 566 patients with traumatic brain injury. The final imputation model included ht ($p<0.001$) and gender ($p=0.01$), with R^2 of 0.97

Steyerbeg. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer in 2009.

Imputation

- **Sample random normal values**
  - Only external information is used
- **Conditional mean** with a single imputation
  - Predictor data only is used
- **Single imputation** with a random draw from the predictive distribution from a imputation model
  - Predictor data and outcome data are used
- **Multiple imputation** with a random draw from the predictive distribution from an imputation model
  - Predictor data and outcome data are used

Imputation

- Problems
  - We may want to predict missing values for one predictor, using other predictors which also have missing values.
- Work around
  - data augmentation methods: which follow an iterative process of an imputation step, which imputes values for the missing data, and a posterior step, which draws new estimates for the model parameters based on the previously imputed values.

Imputation

- Choosing the imputation
  - Imputation model aims to approximate the true distributional relationship between the unobserved data and the available information
  - Two modelling choices usually have to be made:
    - the form of the model (e.g. linear, logistic, polytomous)
    - and the set of variables that enter the model, including potential transformations of predictors.
  - Truncate imputed values, so that they remain within a plausible range
  - Always include all predictors and the outcome of the final model, consider auxiliary predictors.

Imputation

- Multiple Imputation
  - In multiple imputation (MI), missing values are imputed *m* times using *m* independent draws from an imputation model.
  - This means that for each variable with missing data, a conditional distribution for the missing data can be specified given other data
  - *m* completed data sets are created instead of a single completed data set. Missing values are imputed *m* times using *m* independent draws from an imputation model.

Imputation

- Multiple Imputation
  - *m* complete-data analyses are combined to obtain the estimates of regression coefficients and performance estimates
  - As the number of m increases the within variance becomes the stronger overall variance component.
  - The number of m may be as low as 1, when MI becomes single imputation.
  - In prediction research, subjects with missing outcome data are generally discarded.

Imputation

- Steps in dealing with missing data
  - Explore the missing patterns
  - Explore missingness relationship with the outcome
  - Subject matter knowledge should be used to judge plausible mechanisms for the missing values
- Omiting predictors
  - It may be convenient to omit predictors with 50% or more of missigness even if it is of major interest.

Imputation

Table 7.6 Guidelines for reporting of prognostic studies with missing predictor data[61]

Issue	Aspect
Quantification of completeness	If completeness of data is an inclusion criterion, specify numbers excluded Provide total *n* and *n* with complete data Report frequency of missingness for every predictor
Approaches to dealing with missing data	Provide sufficient details on the methods used, including references if imputation was done Specify the *n* of patients and number of events for all analyses
Exploration of missing data	Discuss reasons for missingness Present comparisons of characteristics between cases with and without missing data

Imputation

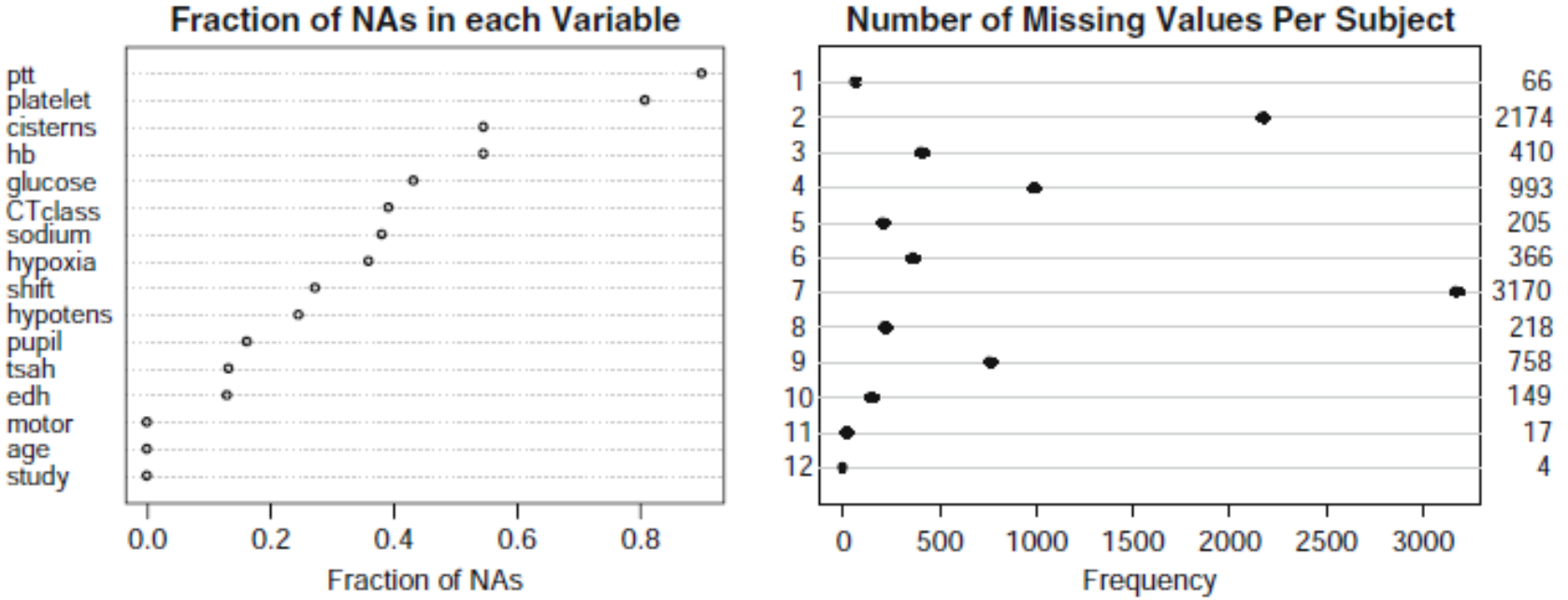

Fig. 8.2 Fraction of missing values per potential predictor (*left panel*), and number of missing values per subject (*right panel*)

Imputation

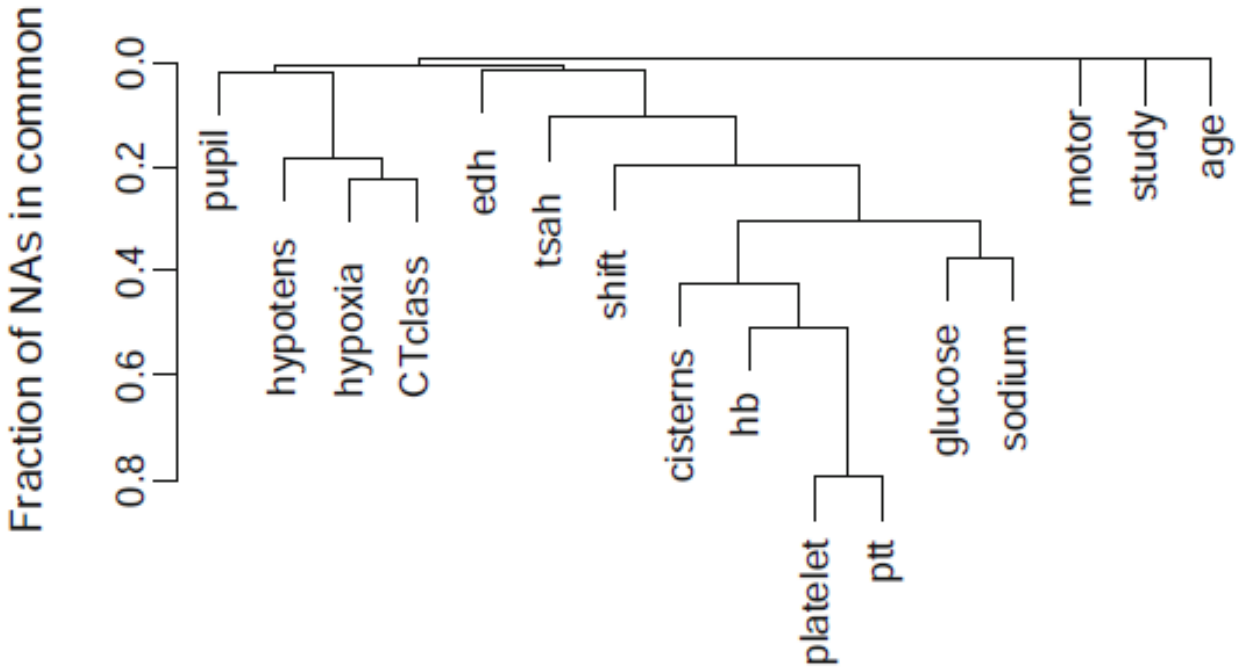

Fig. 8.3 Combinations of missing values in predictors ("NAs"), based on a hierarchical cluster analysis of missingness combinations

Imputation

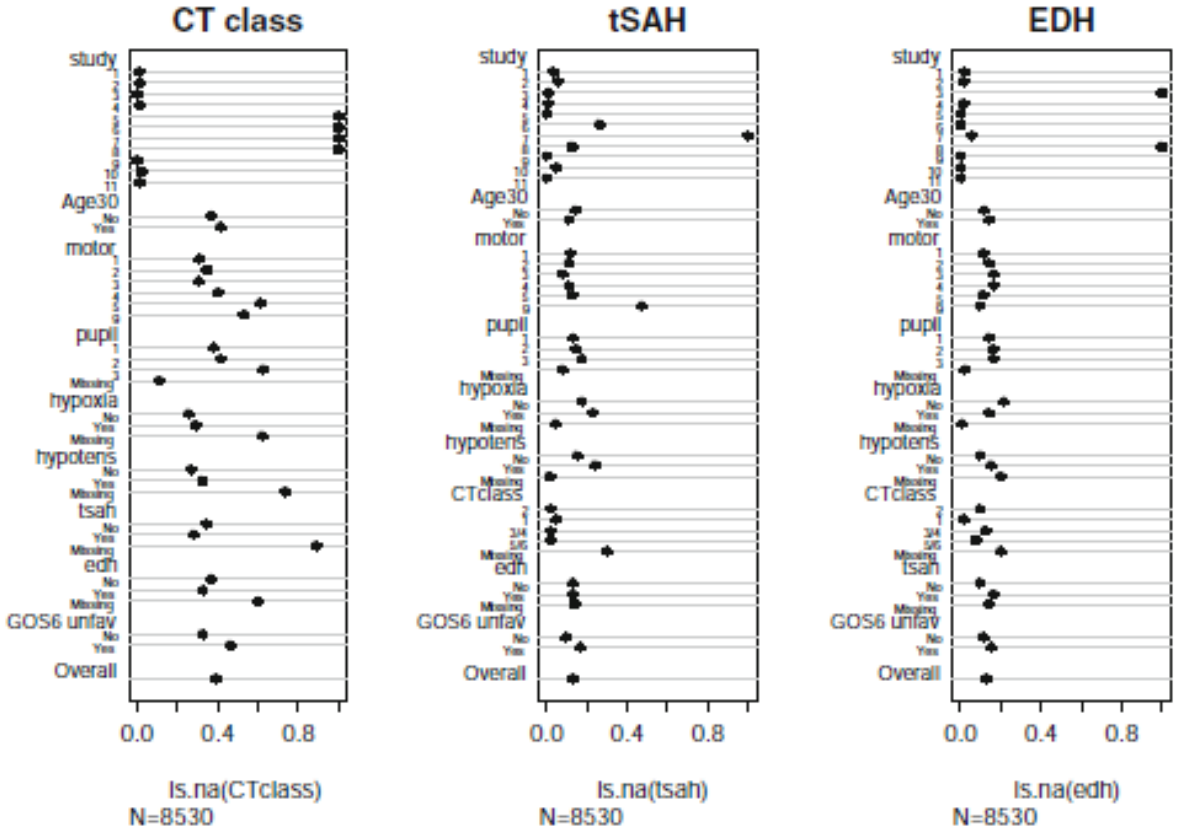

Fig. 8.4 Missingness in relation to study (numbered 1–11), other predictors (age to EDH), and outcome (GOS6). Study was the main determinant of missingness. Only weak associations were observed with other predictors, and no relationship with the 6-month outcome (GOS6)

Conclusions

- With software availability and current evidence pointing the benefits of imputations, it is considered bad practice not to impute data.
- Nevertheless, some analysis steps are not possible with multiple imputed data (e.g. bootstrap optimism estimation) and one must choose a single complete dataset.
- Imputation examples will be shown in the workshop.
- Further reading in multiple imputation are available:
  - https://www.crcpress.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781138588318

fim

Session 11
Dealing with missing data

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br
2018