Missing Data

https://github.com/martingerdin/missing-data-presentation

Martin Gerdin Wärnberg

2022-03-21

Reviewer comment

As the authors state, half the patients were excluded as a result of missing data. Did these patients differ from the patients who were included with regard to age, gender mechanism of injury, severity of TBI, etc?

What does the reviewer want?

Table 1: Comparison of complete and incomplete observations

	Level	Complete	Incomplete
n		8478	7522
Age (median [IQR])		30 [20, 45]	30 [18, 45]
Sex (%)	Female	1627 (19)	2001 (27)
	Male	6851 (81)	5521 (73)
SBP (median [IQR])		116 [100, 128]	110 [100, 124]
RR (median [IQR])		18 [16, 22]	22 [20, 24]
GCS (median [IQR])		15 [9, 15]	14 [8, 15]
Died (%)	No	6735 (79)	5626 (75)
	Yes	1743 (21)	1895 (25)

Abbreviations: GCS Glasgow Coma Scale, RR Respiratory Rate, SBP Systolic Blood Pressure

What is missing data?

- Data that was never collected
- Data that was not available
- Common in all study types

Why is missing data problematic?

- Lower statistical power
- Loss of key subgroups
- Biased or inaccurate estimates
- Increased analysis complexity

What can be done about missing data?

- Avoid it
- Manage it appropriately
- Conduct sensitivity analyses

What are the different missing data mechanisms?

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

What is MCAR?

> ***Observations of all subjects are equally likely to be missing****. That is, there are no systematic differences between subjects with observed and unobserved values meaning that the observed values can be treated as a random sample of the population. For example, echocardiographic measurements might be missing due to sporadic ultrasound malfunction.*
> *– (Papageorgiou et al. 2018)*

What is MAR?

> ***The likelihood of a value to be missing depends on other, observed variables**. Hence, any systematic difference between missing and observed values can be attributed to observed data. That is, the relationships observed in the data at hand can be utilized to 'recover' the missing data. For example, missing echocardiographic measurements might be more normal than the observed ones because younger patients are more likely to miss an appointment.*
>
> *– (Papageorgiou et al. 2018)*

What is MNAR?

> ***The likelihood to be missing depends on the (unobserved) value itself****, and thus, systematic differences between the missing and the observed values remain, even after accounting for all other available information. In other words, there is extra information associated with the missing data that cannot be recovered by utilizing the relationships observed in the data. For example, missing echocardiographic measurements might be worse than the observed ones because patients with severe valve disease are more likely to miss a clinic visit because they are unable to visit the hospital.*
>
> *– (Papageorgiou et al. 2018)*

How can we tell different mechanisms apart?

- We can't use only the data
- Requires knowledge and reasoning about how the data was generated

How can missing data be managed?

- Ignored
- Complete case analysis
- Mean imputation
- Single or deterministic imputation
- Stochastic imputation
- Multiple imputation
- Sensitivity analyses (subgroups, best-worst case scenarios)
- Other ways

How are missing data mechanisms and methods related?

	MCAR	MAR	MNAR
Ignore	Loss of power	Bad idea	Bad idea
Complete case analysis	Loss of power	Biased	Biased
Mean imputation	Biased	Biased	Biased
Single imputation	Biased	Biased	Biased
Stochastic imputation	Probably okay	Biased	Biased
Multiple imputation	Okay	Okay	Probably biased
Sensitivity analyses	Not needed	Good idea	Good idea

Why is ignoring missing data a bad idea?

Treating incomplete data as complete

- Loss of control
- Loss of denominator
- Incomparable groups
- Unknown precision

How can we interpret results based on ignored missing data?

Table 3: Unadjusted and adjusted associations between variables and mortality

	n[a]	Unadjusted OR (95% CI)	Adjusted OR (95% CI)[b]
Age	15978	1.019 (1.017-1.021)	1.025 (1.022-1.029)
Sex	15999	0.804 (0.738-0.876)	0.8 (0.683-0.939)
SBP	12744	0.983 (0.981-0.985)	0.982 (0.979-0.984)
RR	9362	0.973 (0.964-0.982)	1.008 (0.996-1.019)
GCS	13873	0.736 (0.728-0.745)	0.75 (0.739-0.761)

Abbreviations: GCS Glasgow Coma Scale, OR Odds Ratio, RR Respiratory Rate, SBP Systolic Blood Pressure

Notes: [a]The number of complete observations per variable. [b]The number of complete cases used in the adjusted model was 8478.

How can we interpret results based on complete data (complete case analysis)?

Table 4: Unadjusted and adjusted associations between variables and mortality (n = 8478)

	Unadjusted OR (95% CI)	Adjusted OR (95% CI)
Age	1.02 (1.017-1.023)	1.025 (1.022-1.029)
Sex	0.91 (0.799-1.039)	0.8 (0.683-0.939)
SBP	0.981 (0.979-0.983)	0.982 (0.979-0.984)
RR	0.967 (0.957-0.977)	1.008 (0.996-1.019)
GCS	0.748 (0.737-0.758)	0.75 (0.739-0.761)

How is missing data represented in different software?

Software	Missing data representation
Google Sheets (Excel?)	
STATA	., .a-.z
SPSS	.
R	NA

How do software deal with missing data?

- Google Sheets (Excel?), STATA and SPSS will in most cases ignore missing data and use only the observed data to calculate some metric.
- R will in most cases return NA if you try to calculate some metric using data that includes missing values.
- For example, given this vector of systolic blood pressures: `120, 90, 90, NA, 110`, to calculate a mean Google Sheets, STATA, and SPSS would return `102.5` whereas R would return `NA`.
- When doing some regression, STATA, SPSS, and R all default to a complete case analysis.

What is mean imputation?

- Missing values are replaced with the mean of the observed data, for example `120, 90, 90, NA, 110` would become `120, 90, 90, 102.5, 110`.n
- Artificial reduction in variability.

What is single or deterministic imputation?

- Missing values are replaced with the predicted scores from a regression equation.

What is single or deterministic imputation?

- Missing values are replaced with the predicted scores from a regression equation.

Age	Sex	SBP	RR	GCS	Died
40	Male	120	22	15	0
21	Male	110	NA	11	0
3	Female	100	26	15	0
27	Male	130	15	5	0
45	Male	110	NA	15	0
20	Female	100	28	6	1

What is single or deterministic imputation?

- Missing values are replaced with the predicted scores from a regression equation.

Age	Sex	SBP	RR	GCS	Died
40	Male	120	22	15	0
21	Male	110	NA	11	0
3	Female	100	26	15	0
27	Male	130	15	5	0
45	Male	110	NA	15	0
20	Female	100	28	6	1

- Impute RR as:

$$\bar{R}R = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 SBP + \beta_4 GCS + \beta_5 Died$$

What is single or deterministic imputation?

- Running that linear regression results in:

$$\bar{R}R = 15.78 + -0.02 \cdot Age + -0.57 \cdot Sex + 0.03 \cdot SBP + 0.08 \cdot GCS + -0.05 \cdot Died$$

What is single or deterministic imputation?

- Running that linear regression results in:

$$\bar{RR} = 15.78 + -0.02 \cdot Age + -0.57 \cdot Sex + 0.03 \cdot SBP + 0.08 \cdot GCS + -0.05 \cdot Died$$

- Meaning that for this patient:

Age	Sex	SBP	RR	GCS	Died
21	Male	110	NA	11	0

What is single or deterministic imputation?

- Running that linear regression results in:

$$\bar{R}R = 15.78 + -0.02 \cdot Age + -0.57 \cdot Sex + 0.03 \cdot SBP + 0.08 \cdot GCS + -0.05 \cdot Died$$

- Meaning that for this patient:

Age	Sex	SBP	RR	GCS	Died
21	Male	110	NA	11	0

- We could estimate the missing RR as:

$$\bar{R}R = 15.78 + -0.02 \cdot 21 + -0.57 \cdot 1 + 0.03 \cdot 110 + 0.08 \cdot 11 + -0.05 \cdot 0 \approx 19$$

What is stochastic imputation?

- Improves on the single imputation by adding random noise.
  - Single imputation too perfect, amplifies associations (even under MCAR).

What is stochastic imputation?

- Improves on the single imputation by adding random noise.
  - Single imputation too perfect, amplifies associations (even under MCAR).
- The modified regression equation: $\bar{R}R = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 SBP + \beta_4 GCS + \beta_5 Died + \epsilon$
- Where ϵ is the noise term.

What is multiple imputation?

> ***The purpose of multiple imputation is to generate possible values for missing values, thus creating several "complete" sets of data****. Analytic procedures that work with multiple imputation datasets produce output for each "complete" dataset, plus pooled output that estimates what the results would have been if the original dataset had no missing values. These pooled results are generally more accurate than those provided by single imputation methods.*
> *– (SPSS 2021)*

What is multiple imputation?

- While less perfect than single imputation, stochastic imputation still does not incorporate the uncertainty associated with missing data.
- Multiple imputation uses stochastic imputation iteratively.
- Creates multiple complete datasets.
- Each imputed dataset is analysed using standard methods.
- The results from all analyses are pooled to create a combined estimate.

What is multiple imputation?

- You need to decide:
  - How many imputed datasets to create (M).
  - How each variable with missing data should be imputed.
  - What variables to include in the imputation model.
- The defaults:
  - 5 imputed datasets ($M = 5$). Some implementations use 20 ($M = 20$)
  - Logistic regression for binary variables, linear regression for continuous variables.
- More missing data, greater uncertainty, increase M (maybe same as % missingness).

How is multiple imputation implemented?

- STATA: `mi`
- SPSS: `Analyze > Multiple Imputation > Impute Missing Data Values...`
- R: `mice`

Some practical advice

- Unless technically inclined, use complete case analysis.
- Document the amount of missing values in each variable of interest (table).
- Document the number of observations (patients) before incomplete. cases were removed, and the number of observations remaining.
- Show how you went from the original data to the study sample (flowchart).
- Present missing data in the first paragraph in Results.
- Include a table (supplementary material?) comparing complete and incomplete observations.
- Acknowledge possible biases and reflect on these in Limitations.

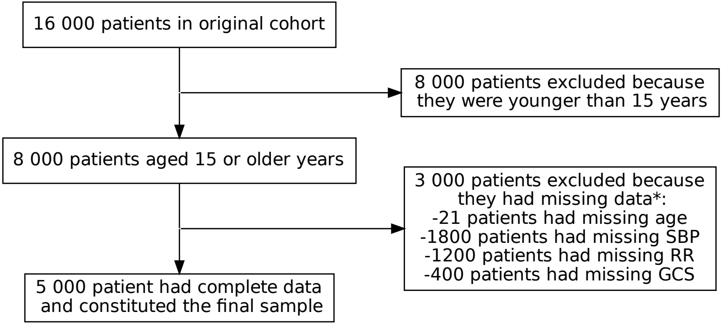
Missing values table

Table 8: Missing values per variable of interest

	n	%
Age	21	0.1
Sex	0	0.0
SBP	3255	20.3
RR	6637	41.5
GCS	2126	13.3
Died	1	0.0

Abbreviations: GCS Glasgow Coma Scale, RR Respiratory Rate, SBP Systolic Blood Pressure

Flowchart



16 000 patients in original cohort

8 000 patients excluded because they were younger than 15 years

8 000 patients aged 15 or older years

3 000 patients excluded because they had missing data*:
-21 patients had missing age
-1800 patients had missing SBP
-1200 patients had missing RR
-400 patients had missing GCS

5 000 patient had complete data and constituted the final sample

Figure 1: Study flowchart

Note: *The sum of missing values may exceed the total number of patients with incomplete data because some patients have missing values in multiple variables.

Table comparing complete and incomplete observations

Table 9: Comparison of complete and incomplete observations

	Level	Complete	Incomplete
n		8478	7522
Age (median [IQR])		30 [20, 45]	30 [18, 45]
Sex (%)	Female	1627 (19)	2001 (27)
	Male	6851 (81)	5521 (73)
SBP (median [IQR])		116 [100, 128]	110 [100, 124]
RR (median [IQR])		18 [16, 22]	22 [20, 24]
GCS (median [IQR])		15 [9, 15]	14 [8, 15]
Died (%)	No	6735 (79)	5626 (75)
	Yes	1743 (21)	1895 (25)

Abbreviations: GCS Glasgow Coma Scale, RR Respiratory Rate, SBP Systolic Blood Pressure

References

Papageorgiou, Grigorios, Stuart W Grant, Johanna J M Takkenberg, and Mostafa M Mokhles. 2018. "Statistical Primer: How to Deal with Missing Data in Scientific Research?†." *Interactive CardioVascular and Thoracic Surgery* 27 (2): 153–58. https://doi.org/10.1093/icvts/ivy102.

SPSS. 2021. "Multiple Imputation." https://www.ibm.com/docs/en/spss-statistics/28.0.0?topic=values-multiple-imputation.