Missing Data

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
- Single imputation
- Multiple imputation
- Sensitivity analyses (subgroups, best-worst case scenarios)
- Other ways

How are missing data mechanisms and methods related?

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

Why is ignoring missing data always a bad idea?

Treating incomplete data as complete

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

Can we interpret results based on ignored missing data?

Table 3: Unadjusted and adjusted associations between variables and mortality

	Unadjusted OR (95% CI)	Adjusted OR (95% CI)
Age	1.019 (1.017-1.021)	1.025 (1.022-1.029)
Sex	0.804 (0.738-0.876)	0.8 (0.683-0.939)
SBP	0.983 (0.981-0.985)	0.982 (0.979-0.984)
RR	0.973 (0.964-0.982)	1.008 (0.996-1.019)
GCS	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

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`.

What is multiple imputation?

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.