Sensitivity of mortality rates to the imputation of missing socioeconomic data: cohort study

DOCUMENT: SAR-2023-017-BH-v01

From: Felipe Figueiredo **To:** Brennan Hickson

2023-12-18

TABLE OF CONTENTS

Sensitivity of mortality rates to the imputation of missing socioeconomic data: cohort study

Document version

Version	Alterations
01	Initial version

1 ABBREVIATIONS

- FIM: Functional Independence Measure
- CI: confidence interval
- DCI: Distress community index
- HR: hazards ratio
- LOCF: Last observation carried forward
- NOCB: Next observation carried backward
- SD: standard deviation
- SES: socioeconomic status
- TBI: Traumatic brain injury

2 CONTEXT

2.1 Objectives

1. To describe the missingness in zip codes at each follow up collection;
2. To impute missing Zip codes with data available in previous follow up collections.
3. To assess the sensitivity of the association between mortality and socioeconomic status to the imputation of participant missing location.

3 METHODS

The data procedures, design and analysis methods used in this report are fully described in the annex document **SAP-2023-017-BH-v01**.

This analysis was performed using statistical software R version 4.3.0.

4 RESULTS

4.1 Missing values in the original dataset

Table 1 shows how the number of outcome events increase when missing zip codes are imputed under the various imputation approaches employed. In all datasets the ID pool remains unchanged, meaning no participant that was dropped was recovered after the imputation approaches evaluated. Only a small number of new outcome events are gained when the multiple observations per individuals are considered. The distribution of frequencies of the SES categories vary slightly between approaches, but these proportions appear to be robust to changes in the underlying zip code imputations.

Table 1 *Distribution of variables in the data under various imputation approaches.*

	Single observation per individual			Multiple observations per individual		
Characteristic	**CC**, N = 7,978[1]	**LOCF**, N = 7,978[1]	**LOCF+NOCB**, N = 7,978[1]	**CC**, N = 24,282[1]	**LOCF**, N = 24,282[1]	**LOCF+NOCB**, N = 24,282[1]
outcome, n	1,003	1,003	1,003	1,006	1,006	1,006
exposure, n (%)						
Prosperous	1,421 (22%)	1,428 (22%)	1,430 (22%)	4,363 (23%)	4,573 (23%)	4,577 (23%)
Comfortable	1,327 (20%)	1,339 (20%)	1,341 (20%)	3,862 (20%)	4,070 (20%)	4,072 (20%)
Mid-Tier	1,221 (19%)	1,237 (19%)	1,238 (19%)	3,573 (19%)	3,764 (19%)	3,765 (19%)
At-Risk	1,287 (20%)	1,296 (20%)	1,299 (20%)	3,782 (20%)	3,993 (20%)	3,997 (20%)
Distressed	1,285 (20%)	1,291 (20%)	1,294 (20%)	3,696 (19%)	3,905 (19%)	3,910 (19%)
Missing	1,437	1,387	1,376	5,006	3,977	3,961

[1]CC = complete-cases (not imputed), LOCF = last observation carried forward, NOCB = next observation carried backward

After dropping incomplete cases to inspect the data available to the model, we can anticipate how the model might be impacted by the imputations (Table 2). Surprisingly all three datasets under the “single observation per individual” approach are the same, and this was validated by the `all.equal()` function that performs a binary comparison between data frames. No changes to the data available for modelling can be detected using any of the imputation approaches for this dataset.

Table 2 *Number of death events available to models in each dataset*

dataset	n
sing_cc	693
sing_locf	693
sing_locf+nocb	693
mult_cc	2
mult_locf	694
mult_locf+nocb	694

Under the “multiple observations per individual” approach, most outcome events are dropped for the complete case dataset. This happens because that dataset uses the exposure at all follow-up times, but most individuals that have expired did not have Zip codes recorded for that follow-up session. This way no DCI scores were available for most individuals, resulting in a sample of size 2 (Table 2). By applying the binary comparison between the two imputation approaches we found that both LOCF and LOCF+NOCB data frames are equal. A single outcome event was added to those datasets after the imputation is applied to the underlying SES data.

This leaves only two datasets to perform the sensitivity analysis on: one dataset under the “single observations per individual” (regardless of whether an imputation was applied) and one using the “multiple observations per individual” approaches (using any imputation). For simplicity, we will consider the complete case dataset for the first case and the LOCF for the second one.

4.2 Sensitivity of proportional hazards violations under different dataset regimens

Table 3 shows the results of the model specification from **SAP-2023-016-BH-v02** on both datasets available from the previous section. The same model specification was tested with both datasets.

Table 3 *Model coefficients for both datasets.*

	Single CC*			Multiple LOCF		
Characteristic	**HR**[1,2]	**95% CI**[2]	**p-value**	**HR**[2]	**95% CI**[2]	**p-value**
SES quintiles						
Prosperous	—	—		—	—	
Comfortable	0.98	0.78 to 1.25	0.893	1.06	0.83 to 1.35	0.623
Mid-Tier	1.09	0.84 to 1.41	0.515	1.18	0.91 to 1.52	0.207
At-Risk	1.12	0.87 to 1.43	0.386	1.11	0.87 to 1.43	0.400
Distressed	1.21	0.95 to 1.56	0.129	1.33	1.03 to 1.72	**0.027**

[1]* FIM scores violate PH assumption

[2]CC = complete-cases (not imputed), CI = Confidence Interval, LOCF = last observation carried forward

Using the dataset that provides a single observation per individual the residual analysis of **SAR-2023-016-BH** is reproduced, where the FIM motor score is dropped due to a violation of the proportional hazards assumption. The dataset that provides multiple observations per individual imputed with a LOCF approach does not violate that assumption, so the term can be safely kept for the analysis. Additionally, when the SES exposure is the time-varying it is associated with mortality under the final model specification, whereas in the smaller dataset this was only true without including any of the FIM scores. Table 4 shows the p-values of the Schoenfeld test for the model tested on both datasets.

Table 4 *Schoenfeld test for both datasets.*

term	cc	locf
exposure	0.5	0.5
SexF	0.2	0.2
Race	0.3	0.4
AGE	0.8	0.8
EDUCATION	>0.9	0.7
EMPLOYMENT	0.3	0.2
RehabPay1	0.6	0.7
SCI	0.2	0.9
DAYStoREHABdc	0.055	0.2
PROBLEMUse	0.4	0.8
ResDis	0.4	0.6
RURALdc	0.4	0.3
FIMMOTD4	0.047	0.055
FIMCOGD4	0.2	0.13
GLOBAL	0.3	0.2

5 OBSERVATIONS AND LIMITATIONS

Recommended reporting guideline

The adoption of the EQUATOR network (http://www.equator-network.org/) reporting guidelines have seen increasing adoption by scientific journals. All observational studies are recommended to be reported following the STROBE guideline (von Elm et al, 2014).

6 CONCLUSIONS

Simple imputation on zip codes do not affect the range of observations available for modeling in this dataset when using a single observation per individual. The model specification tested is robust to imputation approaches on this dataset and the resulting exposure variable is unchanged.

When using multiple observations per individual, there is a minute increment in the number of events, but there is different imputation approaches do not yield different datasets. The model specification tested is robust to imputation approaches on this dataset and the resulting exposure variable is unchanged.

When using multiple observations per individual in the model specification evaluated, the time-varying exposure allows for the inclusion of terms that violated the proportional hazards assumption in the constant exposure. The model specification tested is sensitive to using a time-varying exposure and all terms can be used for analysis.

7 REFERENCES

- **SAP-2023-017-BH-v02** – Analytical Plan for Sensitivity of mortality rates to the imputation of missing socioeconomic data: cohort study
- **SAP-2023-016-BH-v02** – Analytical Plan for Time-adjusted effect of socioeconomic status in mortality rates after brain injury: cohort study
- von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg. 2014 Dec;12(12):1495-9 (https://doi.org/10.1016/j.ijsu.2014.07.013).

8 APPENDIX

8.1 Exploratory data analysis

N/A

8.2 Availability

All documents from this consultation were included in the consultant's Portfolio.

The portfolio is available at:

https://philsf-biostat.github.io/SAR-2023-017-BH/

8.3 Associated analyses

This analysis is part of a larger project and is supported by other analyses, linked below.

Effect of socioeconomic status in mortality rates after brain injury: cohort study

https://philsf-biostat.github.io/SAR-2023-004-BH/

Time-adjusted effect of socioeconomic status in mortality rates after brain injury: cohort study

https://philsf-biostat.github.io/SAR-2023-016-BH/

8.4 Analytical dataset

Table A1 shows the structure of the analytical dataset.

Table A1 *Analytical dataset structure*

id	exposure	outcome	Time	SexF	Race	Mar	AGE	PROBLEMUse	EDUCATION	EMPLOYMENT	RURALdc	PriorSeiz	SCI	Cause	RehabPay1	ResDis	DAYStoREHABdc	FIMMOTD	FIMCOGD	FollowUpPeriod	FIMMOTD4	FIMCOGD4
1																						
2																						
3																						
…																						
N																						

Due to confidentiality the data-set used in this analysis cannot be shared online in the public version of this report.