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

DOCUMENT: SAP-2023-017-BH-v02

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

2023-07-19

TABLE OF CONTENTS

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

Document version

Version	Alterations
01	Initial version
02	added full datasets for time-varying SES values

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.

2.2 Hypotheses

1. Imputing the missing zip codes will decrease the missingness in the dataset and improve the model fit.
2. Switching to a time-varying covariates approach will allow for the addition of covariates that violated the proportional hazards assumption

3 DATA

3.1 Raw data

The raw data table was created by merging the TBI database with the DCI table, using the Zip codes as merging key. The raw data base had 711 variables collected with 76,665 observations from 19,303 individuals.

From the raw table, multiple analytical datasets will be created by applying various imputation methods to the Zip code values. The creation of the analytical datasets is described in the next section and the imputation procedures are described in section 5.1.4.

3.2 Analytical dataset

Many datasets will be created for this sensitivity analysis, and the many-datasets approach will be used to apply the statistical model (defined in section 5.1.3) to each dataset so that a sensitivity analysis can be performed. The datasets created under this approach will be created in steps, and stored in a single object to which specific code instructions can be applied to all datasets in a single command. This approach will allow for the simultaneous application of the following instructions to all datasets:

1. all data cleaning procedures
2. inclusion/exclusion criteria
3. the statistical model selected for evaluation
4. the calculation of model performance metrics

After the cleaning process 24 variables were included in the analysis. The total number of observations excluded due to incompleteness and exclusion criteria will be reported in the analysis. Table 1 shows the structure of the analytical dataset.

Table 1 *Analytical dataset structure after variable selection and cleaning.*

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																						

All variables in the analytical set were labeled according to the raw data provided and values were labeled according to the data dictionary for the preparation of production-quality results tables and figures.

4 STUDY PARAMETERS

4.1 Study design

This is a retrospective analysis of a prospective cohort study.

4.2 Inclusion and exclusion criteria

Inclusion criteria

1. Participants with at most 10 years of follow up;
2. Participants included in the cohort between 2010-01-01 and 2018-12-31;
3. Only the last valid observation of each individual will be included in the analysis*.

* this criterion will be treated specially in this analysis (see section 5.1.4).

Exclusion criteria

1. Observations after 2019-12-31 will be excluded in order to mitigate risk of confounding by COVID-19 related deaths.
2. Observations prior to this date will still be considered for participants where such data is available.

4.3 Exposures

SES of the neighborhood to which the participant was discharged. The SES measure was stratified into its quintiles, and labelled according to the data dictionary to facilitate interpretation of the results.

4.4 Outcomes

Specification of outcome measures (Zarin, 2011):

1. (Domain) Mortality
2. (Specific measurement) Death
3. (Specific metric) Time-to-event
4. (Method of aggregation) Hazard ratio

Primary outcome

Death after a brain injury.

4.5 Covariates

- Sex
- Race
- Age at injury

- Substance Problem Use
- Education
- Employment status
- Rural area
- Previous seizure disorder diagnosis
- Spinal cord injury
- Cause of injury
- Primary rehabilitation payer
- Residence after rehab discharge
- Days From Injury to Rehab Discharge
- FIM Motor at Discharge
- FIM Cognitive at Discharge

5 STATISTICAL METHODS

5.1 Statistical analyses

5.1.1 Descriptive analyses

The epidemiological profile of the study participants will be described. Demographic and clinical variables will be described as mean (SD) or as counts and proportions (%), as appropriate. The distributions of participants' characteristics will be summarized in tables and visualized in exploratory plots.

5.1.2 Inferential analyses

All inferential analyses will be performed in the statistical models (described in the next section).

5.1.3 Statistical modeling

This analysis will evaluate the sensitivity to the model specification chosen in **SAR-2023-016-BH** to changes in the SES data (defined as the exposure of that analysis). The model specification used for the sensitivity analysis will be the best model selected in that associated report.

For reference, the specification defined there regresses the hazard on the SES controlling for all covariates listed in section 4.5, except "previous seizure".

This model specification will be applied on all datasets created from the imputation approaches described in section 5.1.4, and the Schoenfeld test will be applied to verify the proportional hazards assumption on all model terms.

5.1.4 Missing data

A couple of simple imputation approaches will be applied on missing values for Zip codes, before the DCI data is merged into the TBI database. An LOCF-based imputation will be applied to impute future Zip codes based on the last known value for each individual. An additional dataset will be created by applying both NOCB- and LOCF-based imputations on missing values, with the intention of increasing the proportion of location data before the DCI data is merged and inclusion/exclusion criteria are applied, in particular the criterion that selects only the last valid observation of the individual in the study period. The non-imputed complete case dataset will be used as the control for the evaluation of the LOCF and the NOCB+LOCF datasets.

In addition to the datasets created as described above, a separate set of tables will be created without the application of inclusion criterion #3 (section 4.2). Thus a new batch of datasets will be tested that include multiple observations per individual to assess model performance under time-varying SES values.

5.2 Significance and Confidence Intervals

All analyses will be performed using the significance level of 5%. All significance hypothesis tests and confidence intervals computed will be two-tailed.

5.3 Study size and Power

N/A

5.4 Statistical packages

This analysis will be performed using statistical software R version 4.3.0.

6 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).

7 REFERENCES

- **SAR-2023-017-BH-v01** – Sensitivity of mortality rates to the imputation of missing socioeconomic data: cohort study
- **SAR-2023-016-BH** – Time-adjusted effect of socioeconomic status in mortality rates after brain injury: cohort study
- Zarin DA, et al. The ClinicalTrials.gov results database – update and key issues. N Engl J Med 2011;364:852-60 (https://doi.org/10.1056/NEJMsa1012065).
- Gamble C, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. JAMA. 2017;318(23):2337–2343 (https://doi.org/10.1001/jama.2017.18556).
- 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

This document was elaborated following recommendations on the structure for Statistical Analysis Plans (Gamble, 2017) for better transparency and clarity.

8.1 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/

8.2 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/