Handling missing data and errors in Estonian eHealth information system

Viktoria Kirpu

National Institute for Health Development of Estonia

August 2018

Co-author: *MSc* Natalja Eigo

Collecting data

Health care providers

EHIF

eHIS

NIHD

Target:
eHIS

eHIS and EHIF

DIGILUGU
EESTI E-TERVISE SIHTASUTUS

eHIS

- Database, a part of the state health information system
- Health care providers oblige to provide epicrisis and other medical documents
- Used for keeping records of state of health
- Data used for producing health statistics

Eesti
Haigekassa

EHIF

- Task is to organize health insurance in order to enable health insurance benefits for insured persons
- Collects documentation about invoices for treatment cases from facilities providing health care services

Problems caused by non-response

- Loss of the necessary information
- Reduction of the capacity of the study
- Bias in the estimates assessments

The smaller the bias,
the better statistical results reflect the actual situation.

Data set **without non-response and errors**

$U = \{1, ..., k, ..., N\}$ – population of the size N

y_k – value of variable Y

The total sum of variable Y:

$$Y = y_1 + ... + y_N = \sum_{k=1}^{N} y_k.$$

Data set **with non-response**

There is almost always a non-response in empirical data.

Types of data with missing values:

- *unit non-response* – observation is missing
- *item non-response* – only part of the response is missing

Data set **with errors** (1)

Types of errors in eHIS:

- *random errors* – caused by inaccuracy of measuring or recording
- *systematic errors* – mainly caused by the inaccuracy of the instrument
- *gross errors* – the value of the characteristic is outside the area of possible values for the characteristic
- *logical errors* - where the values of various characteristics are inconsistent

Data set **with errors** (2)

Upon discovery of errors, they must be eliminated and treated as non-response if necessary.

Main methods:

- using additional information
- imputation

Using additional information

NIHD uses EHIF data as an additional source to improve eHIS data.

- **InfoU –**
  Assisting information: vector $x_k = x_k^*$
  $\Leftarrow$ known for each $k \in N$
  Additional information: the total sums vector $X^* = \sum_U x_k^*$
  $\Leftarrow$known on the level of population U
- **InfoS –**
  Assisting information: vector is $x_k = x_k^\circ$
  $\Leftarrow$ known within the existing data but not on the level of the population
- **InfoUS –**
  Assisting information: vector $x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix}$
  $\Leftarrow$ known on the level of the population as well as the sample

Imputation

The imputation methods:

- statistical prediction methods
- getting values from responded similar objects and replacing for those who have not responded
- expert opinion

Imputed values are artificial - imputed values always differ from the actual values of objects to some extent!

There is no good reason for careful imputation to create more damage to assessments than other methods when producing statistics.

Summary

- Essential to know whether the dataset is complete
- In case of data without losses, we can produce statistics immediately
- In case of data with losses, it is necessary to carry out data processing in advance by using additional information and/or imputation
- Statistics mistakenly produced from data with losses will result biased estimates - doesn't reflect the actual situation
- In the current eHIS data quality control NIHD wishes to use data received from EHIF

Thank You for Your attention! ☺