Handling missing data and errors in Estonian eHealth information system

Viktoria Kirpu[1] and Natalja Eigo[2]

[1]National Institute for Health Development of Estonia, e-mail: viktoria.kirpu@tai.ee
[2]National Institute for Health Development of Estonia, e-mail: natalja.eigo@tai.ee

Abstract

An overview about data loss and errors in Estonian eHealth information system is provided in this paper. Besides, some statistical imputation methods are described to solve these problems. At the end, possibilities of using additional information for improving data are discussed.

Keywords: non-response, loss, errors, imputation, additional information, biased assessment, unbiased assessment

1 Introduction

In Estonia health statistics are collected, processed, analysed and published by National Institute for Health Development (Estonian: *Tervise Arengu Instituut*, also called as NIHD). NIHD uses the eHealth information system or eHIS (Estonian: *Tervise infosüsteem*) as one data source for statistics. Unfortunately, the database concerned has deficiencies in the coverage and quality of the data. To validate eHIS data NIHD uses treatment invoices data provided by Estonian Health Insurance Fund (Estonian: *Eesti Haigekassa*, also called as EHIF). At the present EHIF has received more observations than the eHIS.

However if data has not been sent to the database, it is important to take this fact into account when computing statistics and to implement necessary statistical methods. Otherwise, incomplete data may lead to biased estimates that do not correspond to the population.

1.1 eHIS data

The eHealth information system or eHIS was created in 2008 and is managed and developed now by Health and Welfare Information Systems Centre (Estonian: *Tervise- ja Heaolu Infosüsteemide Keskus*, also called as TEHIK). eHIS an important database which is a part of the state health information system (Estonian eHealth Foundation, n.d.). Health care providers oblige to provide epicrisis and other medical documents to eHIS (Riigi Teataja I, 2018). This system's data among other functionalities is used for keeping records of state of health and for producing health statistics (National Institute for Health Development, 2017).

1.2 Estonian Health Insurance Fund data

The most important task of Estonian Health Insurance Fund is to organise health insurance in order to enable health insurance benefits for insured persons. In addition, the task of Health

Insurance Fund is to assist with preparing standards of treatment and treatment guidelines, motivate health care providers to develop quality of health services, organise the performance of international agreements concerning health insurance; participate in planning of health care. (Estonian Health Insurance Fund, n.d.) Estonian Health Insurance Fund also collects documentation about invoices for treatment cases from facilities providing health care services.

2 Problems related to non-response

Non-response occurs in the analysed database when documentation about treatment case has not been submitted to the Health Information System (eHIS).

Lack of data does not only cause a loss of the necessary information and a reduction of the capacity of the study[1], but it causes bias in the estimates assessments[2]. It is crucial to minimise the number of undiscovered lost observations i.e. number of treatment cases, concerning which documentation was not submitted to eHIS and the lack of which was not discovered during checking. Otherwise, statistical conclusions, for example the confidence interval, may be estimated incorrectly. It is necessary to have an unbiased estimate assessment[3] for high-quality statistics or the bias should be reduced as much as possible. The smaller the bias, the better statistical results reflect the actual situation.

For example, the emergency type of a treatment case is the one that usually is not submitted to eHIS by the doctors. If there is a situation, where doctors do not note down emergency treatment cases, then it gives the impression that there are few that kind of treatment cases in the country. In such a situation, we can be certain that the received statistics do not describe the reality and we have received biased estimates of assessments. In other common case biases also arise when whole epicrisis has not even been provided.

There is a strong believing that if the rate of response is high, it is not important to take into account the non-response. Statistics does not focus on the rate of response as an indicator, which reduces the bias caused by non-response, as the rate of response itself does not measure it. Unlike variance, the bias does not near zero when increasing the sample size (Shouten & Cobben, 2007; Särndal & Lundström, 2005). In order to reduce the bias caused by non-response it is vital to use the necessary methods of assessment.

3 Handling a data set without non-response

Let $U = \{1, ..., k, ..., N\}$ be the population of the size N and y_k value of variable Y. Then the total sum of variable Y is: $Y = y_1 + ... + y_N = \sum_{k=1}^{N} y_k$. (Estonian eHealth Foundation, n.d.) In this case, we can get the value of variable Y as the eHIS data set is complete i.e. the patient epicrisis of all treatment cases have been provided to eHIS and there are characteristic values for all observations.

4 Handling a data set with errors and non-response

There is almost always a non-response in empirical data. Data with missing values occures in two different ways: when the observation is missing (unit non-response) or when only part

[1] An indicator (probability), which assesses how important the received result is for statistical purposes (Aron & Aron, 1997).

[2] An assessment, the mean value of which differs from the true value of the assessed parameter by a certain systematic error.

[3] An assessment, the mean value of which equals the true value of the assessed parameter.

of the response is missing (item non-response) (Särndal & Lundström, 2005; Andridge & Little, 2010). In our case, unit non-response means that the patient epicrisis wasn't submitted to eHIS and item non-response means that the patient epicrisis has been submitted with incomplete information. Errors in eHIS can be divided into the following types:

- **random errors** (caused by inaccuracy of measuring or recording, generally have little effect on the result and are difficult to discover), for example the wrong month of birth of the patient;
- **systematic errors** (mainly caused by the inaccuracy of the instrument), for example when the doctor uses the same diagnosis to describe all illnesses of the patients;
- **gross errors** (the value of the characteristic is outside the area of possible values for the characteristic), for example a cervical cancer diagnosis for a male patient;
- **logical errors**, where the values of various characteristics are inconsistent, for example the date of discharging a person from the hospital is marked to be after the date of death.

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

A data set with non-response cannot be processed in the same way as data with no non-response. The main methods for handling data with non-response is **using additional information** and **imputation**[4].

5 Using additional information

In the case of eHIS data loss, NIHD uses EHIF data as an additional source to improve data.

Three types of additional information are distinguished in case of data sets with loss: *InfoU*, *InfoS*, *InfoUS* (Särndal & Lundström, 2005).

- **InfoU**
  For this type of information the assisting information is vector $x_k = x_k^*$, which is known for each $k \in N$. Additional information, which is the total sums vector $X^* = \sum_U x_k^*$, is known on the level of population U.
- **InfoS**
  For this type of information the assisting information is known within the existing data but not on the level of the population and the assisting vector is $x_k = x_k^\circ$.
- **InfoUS**
  For this type of information $x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix}$ is known on the level of the population as well as the sample.

Values found from an additional source can be used for the replacement of missing data in eHIS. It helps to reduce bias of estimates.

In order to "enrich" eHIS data NIHD uses data received from EHIF as *InfoU* additional information. NIHD presumes that the data in this certain database has no item non-responses. But not all needed information can be found in this data source. For remaining missing data imputation methods should be applied.

[4] A procedure, where the missing values of one or several variables are replaced by the estimates based on the existing data or some other information.

6 Data imputation

The imputation methods of values are divided into three groups:

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

The first two groups are classified as statistical methods. The methods of the first group are based on the relationship between variables such as finding regression models. The methods of the second group can be called donor-based, because the imputed value is borrowed from some observed object, which is similar to the missing object. The methods of the third group largely depend on the expert's skills and knowledge.

Imputation methods can be further classified as deterministic, i.e. when repeating the imputation procedure one always archieve at exactly the same result, or random, where different imputed values are received when repeating the procedure. Regression imputation is an example of a deterministic method. When we however impute the value of a randomly selected similar donor, it is a random method that is often used for Hot-Deck imputation.

It is important to take into account that imputed values are artificial - they are either construed according to some rule or the values of other responded objects. Therefore, imputed values always differ from the actual values of objects to some extent. It is expected that in case of imputation the estimates have a small dispersion and no bias or a small bias.

Generally imputation must be approached with caution as we are trying to produce reliable statistics from data we know is imprecise to a lesser or greater extent from the beginning. At the same time, it is often necessary to do so for practical reasons. There is no good reason for careful imputation to create more damage to assessments than other methods when producing statistics.

7 Summary

In order to produce statistics it is essential to know whether we are using a full dataset to analyze data, or one with losses. Different methods should be implemented for producing statistics according to this prior knowledge. In the case of data without losses, we can produce statistics immediately. In case of data with losses and/or errors, it is necessary to carry out data processing in advance. Various methods are used for this purpose. For example, missing values can be estimated by engaging additional information and imputation. Statistics mistakenly produced from data with losses will result biased estimates. By finding values for missing data, it is possible to reduce the bias or in some cases even eliminate it.

Therefore, in order to produce health statistics the quality of eHIS data must be checked, i.e. whether all necessary documents of treatment cases have been provided to the health information system.

In the current eHIS data quality control NIHD wishes to use data received from EHIF as InfoU additional information. Thanks to this, we would be able to find out how much data is not submitted to the health information system and how many errors are made when submitting data. According to this, it would be possible to use additional information and imputation for supplementing the health information system data set.

References

Andridge, R. R. & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* **78(1)**, 40 — 64.

Aron, A. & Aron, E. N. (1997). *Statistics for the behavioral and social sciences: A brief course.* NJ: Prentice Hall.

Särndal, C. & Lundström, S. (2005). *Estimation in surveys with nonresponse.* John Wiley and Sons: New-York.

Shouten, B. & Cobben, F. (2007). *R-indexes for comparison of different fieldwork strategies and data collection modules.* Voorburg-Heerlen: Statistics Netherland. Discussion paper 07002.

Estonian Health Insurance Fund (n.d.). Haigekassa organisatsioon. Last checked: 06.02.2018. https://www.haigekassa.ee/haigekassa/organisatsioon.

Estonian eHealth Foundation (n.d.). Tervise infosüsteem. Last checked: 06.02.2018. http://www.e-tervis.ee/index.php/et/eesti-etervise-sihtasutus/tervise-infosusteem.

National Institute for Health Development (2017). E-tervise infosüsteem. Last checked: 31.01.2018, http://www.e-tervis.ee/index.php/et/eesti-etervise-sihtasutus/tervise-infosusteem.

Riigi Teataja I (2018). Health Services Organisation Act[1], § 592. Last checked: 09.04.2018. https://www.riigiteataja.ee/akt/TTKS.